Chapter 9

Simple Linear Regression

The relationship between two continuous variables, sediment concentration and stream discharge, is to be investigated. Of interest is the quantification of this relation into a model form for use as a predictive tool during days in which discharge was measured but sediment concentration was not. Some measure of the significance of the relationship is desired so that the analyst can be assured that it is in fact composed of more than just background noise. A measure of the quality of the fit is also desired.

Sediment concentrations in an urban river are investigated to determine if installation of detention ponds throughout the city have decreased instream concentrations. Linear regression is first performed between sediment concentration and river discharge to remove the variation in concentrations which are due to flow variations. After subtracting this linear relation from the data, the residual variation before versus after the installation of ponds can be compared to determine their effect.

Regression of sediment concentration versus stream discharge is performed to obtain the slope coefficient for the relationship. This coefficient is tested to see if it is significantly different than a value obtained 5 years before using a rainfall-runoff model of the basin.

The above examples all perform a linear regression between the same two variables, sediment concentration and water discharge, but for three different objectives. Regression is commonly used for at least these three objectives. This chapter will present the assumptions, computation and applications of linear regression, as well as its limitations and common misapplications by the water resources community.

Ordinary Least Squares (OLS), commonly referred to as linear regression, is a very important tool for the statistical analysis of water resources data. It is used to describe the covariation between some variable of interest and one or more other variables. Regression is performed to

1) learn something about the relationship between the two variables, or
2) remove a portion of the variation in one variable (a portion that is not of interest) in order to gain a better understanding of some other, more interesting, portion of the variation, or
3) estimate or predict values of one variable based on knowledge of another variable, for which more data are available.

This chapter deals with the relationship between one continuous variable of interest, called the **response variable**, and one other variable -- the **explanatory variable**. The name "simple linear regression" is applied because one explanatory variable is the simplest case of regression models. The case of multiple explanatory variables is dealt with in Chapter 11 -- multiple regression.

9.1 The Linear Regression Model

The model for simple linear regression is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i=1,2,....,n$$

where

- y_i is the ith observation of the response (or dependent) variable
- x_i is the ith observation of the explanatory (or independent) variable
- β_0 is the intercept
- β_1 is the slope
- ε_i is the random error or residual for the ith observation, and
- n is the sample size.

The error around the linear model ε_i is a random variable. That is, its magnitude is not controlled by the analyst, but arises from the natural variability inherent in the system. ε_i has a mean of zero, and a constant variance σ^2 which does not depend on x. Due to the latter, ε_i is independent of x_i.

Regression is performed by estimating the unknown true intercept and slope β_0 and β_1 with b_0 and b_1, estimates derived from the data. As an example, in figure 9.1 the true linear relationship between an explanatory variable x and the response variable y is represented by a solid line. Around the line are 10 observed data points which result from that relationship plus the random error ε_i inherent in the natural system and the process of measurement. In practice the true line

is never known -- instead the analyst measures the 10 data points and estimates a linear relationship from those points. The OLS estimate developed from the 10 measurements is shown as the dashed line in figure 9.2.

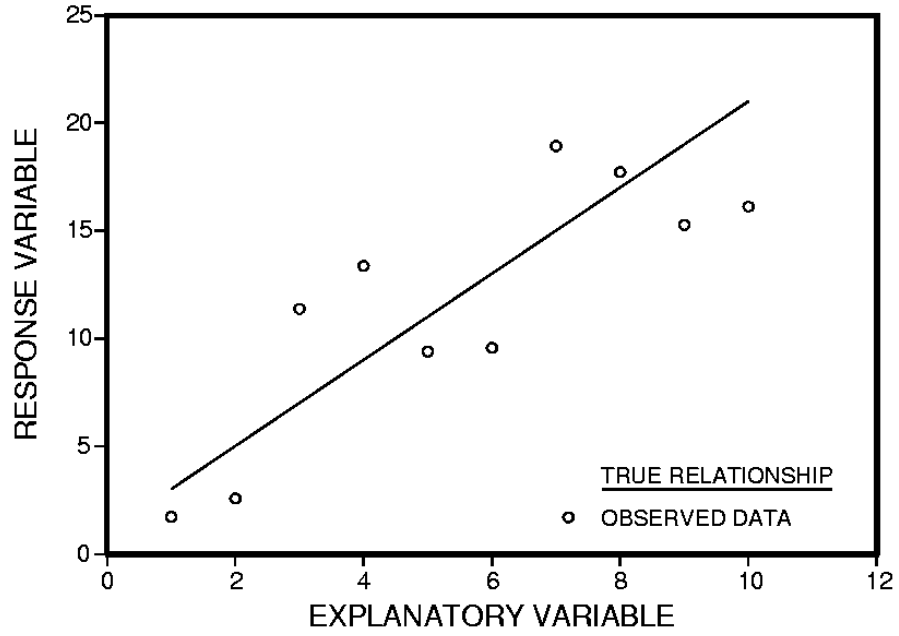


Figure 9.1 True linear relation between x and y, and 10 resultant measurements.

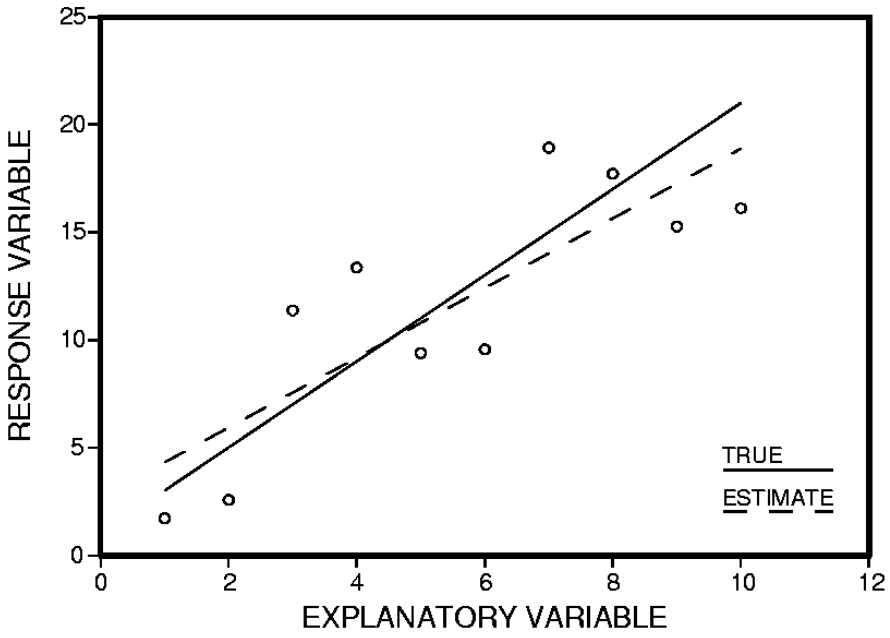


Figure 9.2 True and estimated linear relation between x and y.

If 10 new data points resulting from the same true (solid line) relationship are measured and their OLS line computed, slightly different estimates of b_0 and b_1 result. If the process is repeated several times, the results will look like figure 9.3. Some of the line estimates will fall closer to the true linear relationship than others. Therefore a regression line should always be considered as a sample estimate of the true, unknown linear relationship.

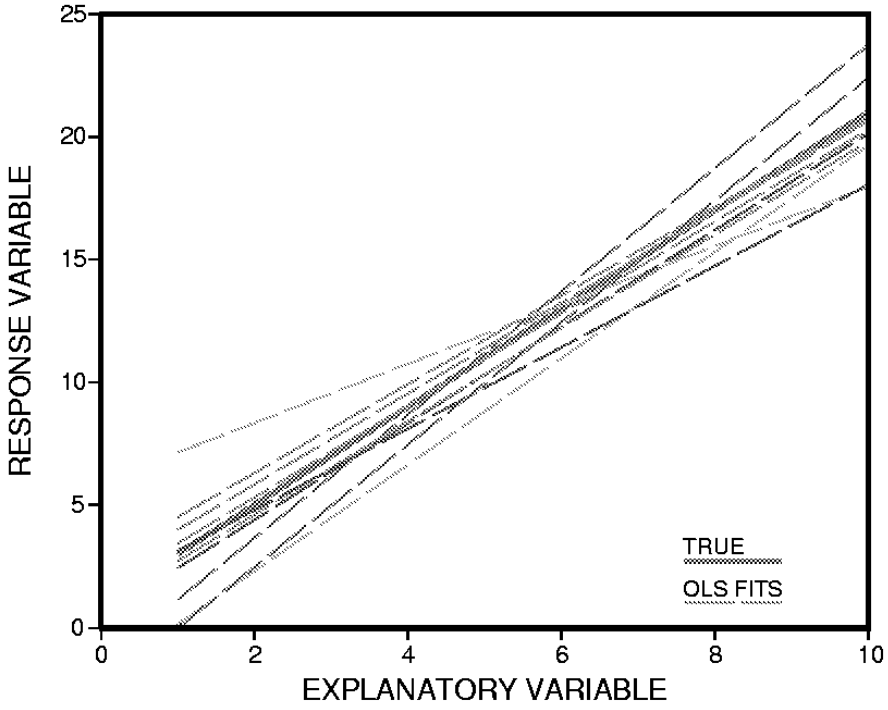


Figure 9.3 True and several estimated linear relations between x and y.

Another way of describing the linear regression model is as an estimate of the mean of y, given some particular value for x. This is called a conditional distribution. If x takes on the value x_0, then y has a conditional mean of $\beta_0 + \beta_1 x_0$ and conditional variance σ^2. The mean is "conditioned", or depends on, that particular value of x. It is the value expected for y given that x equals x_0. Therefore:

the "expected value" of y given x_0	$E[y \mid x_0]$	$= \beta_0 + \beta_1 x_0$
the variance of y given x_0	$Var[y \mid x_0]$	$= \sigma^2$

9.1.1 Assumptions of Linear Regression

There are five assumptions associated with linear regression. These are listed in table 9.1. The necessity of satisfying them is determined by the purpose to be made of the regression equation. Table 9.1 indicates for which purposes each is needed.

	Purpose			
Assumption	Predict y given x	Predict y and a variance for the prediction	Obtain best linear unbiased estimator of y	Test hypotheses, estimate confidence or prediction intervals
(1) Model form is correct: y is linearly related to x	+	+	+	+
(2) Data used to fit the model are representative of data of interest.	+	+	+	+
(3) Variance of the residuals is constant (is homoscedastic). It does not depend on x or on anything else (e.g. time).		+	+	+
(4) The residuals are independent.			+	+
(5) The residuals are normally distributed.				+

Table 9.1 Assumptions necessary for the purposes to which OLS is put.
+: the assumption is required for that purpose.

The assumption of a normal distribution is involved only when testing hypotheses, requiring the residuals from the regression equation to be normally distributed. In this sense OLS is a parametric procedure. No assumptions are made concerning the distributions of either the explanatory or response variables. The most important hypothesis test in regression is whether the slope coefficient is significantly different from zero. Normality of residuals is required for this test, and should be checked by a boxplot or probability plot. The regression line, as a conditional mean, is sensitive to the presence of outliers in much the same way as a sample mean is sensitive to outliers.

9.2 Computations

Linear regression estimation is nothing more than a minimization problem. It can be stated as follows: find two numbers b_0 and b_1 such that

$\sum_{i=1}^{n} (y_i - y_i)^2$ is minimized, where $\hat{y}_1$ is the OLS estimate of y:

$$\hat{y}_i = b_0 + b_1 x_i.$$

This can be solved for b_0 and b_1 using calculus. The solution is referred to as the normal equations. From these come an extensive list of expressions used in regression:

Formula	Name
$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$	mean x
$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$	mean y
$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$	sums of squares y = Total SS
$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$	sums of squares x
$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i y_i) - n\,\bar{x}\,\bar{y}$	sums of x y cross products
$b_1 = S_{xy} / SS_x$	the estimate of β_1 (slope)
$b_0 = \bar{y} - b_1\,\bar{x}$	the estimate of β_0 (intercept)
$\hat{y}_i = b_0 + b_1 x_i$	the estimate of y given x_i

Formula	Name
$e_i = y_i - \hat{y}_i$	the estimated residual for obs. i
$SSE = \sum_{i=1}^{n} e_i^2$	error sum of squares
$s^2 = (SS_y - b_1 S_{xy}) / (n-2)$ $= \sum_{i=1}^{n} e_i^2 / (n-2)$	The estimate of σ^2, also called mean square error (MSE).
$s = \sqrt{s^2}$	standard error of the regression or standard deviation of residuals
$SE(\beta_1) = s / \sqrt{SS_x}$	standard error of β_1
$SE(\beta_0) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$	standard error of β_0
$r = S_{xy} / \sqrt{SS_x SS_y}$ $= b_1 \sqrt{SS_x / SS_y}$	the correlation coefficient
$R^2 = [SS_y - s^2 (n-2)] / SS_y$ $= 1 - (SSE / SS_y)$ $= r^2$	coefficient of determination, or fraction of the variance explained by regression

9.2.1 Properties of Least Squares Solutions

1) If assumptions 1 through 4 are all met, then the estimators b_0 and b_1 are the minimum variance unbiased estimators of β_0 and β_1.
2) The mean of the residuals (e_i's) is exactly zero.
3) The mean of the predictions ($\hat{y}_i$'s) equals the mean of the observed responses (y_i's).
4) The regression line passes through the centroid of the data ($\bar{x}$, $\bar{y}$).
5) The variance of the predictions ($\hat{y}_i$'s) is less than the variance of the observed responses (y_i's) unless $R^2 = 1.0$.

9.3 Building a Good Regression Model

A common first step in performing regression is to plug the data into a statistics software package and evaluate the results using R^2. Values of R^2 close to 1 are often incorrectly deemed an indicator of a good model. This is a dangerous, blind reliance on the computer software. An R^2 near 1 can result from a poor regression model; lower R^2 models may often be preferable. Instead of the above, performing the following steps in order will generally lead to a good regression model.

The following sections will use the total dissolved solids (TDS) concentrations from the Cuyahoga River at Independence, Ohio, 1974-1985 as an example data set. The data are found in Appendix C9. These concentrations will be related to stream discharge (Q).

1) First step -- PLOT THE DATA!
   Plot y versus x and check for two things
   1a) does the relationship look non-linear?
   1b) does the variability of y look markedly different for different levels of x?

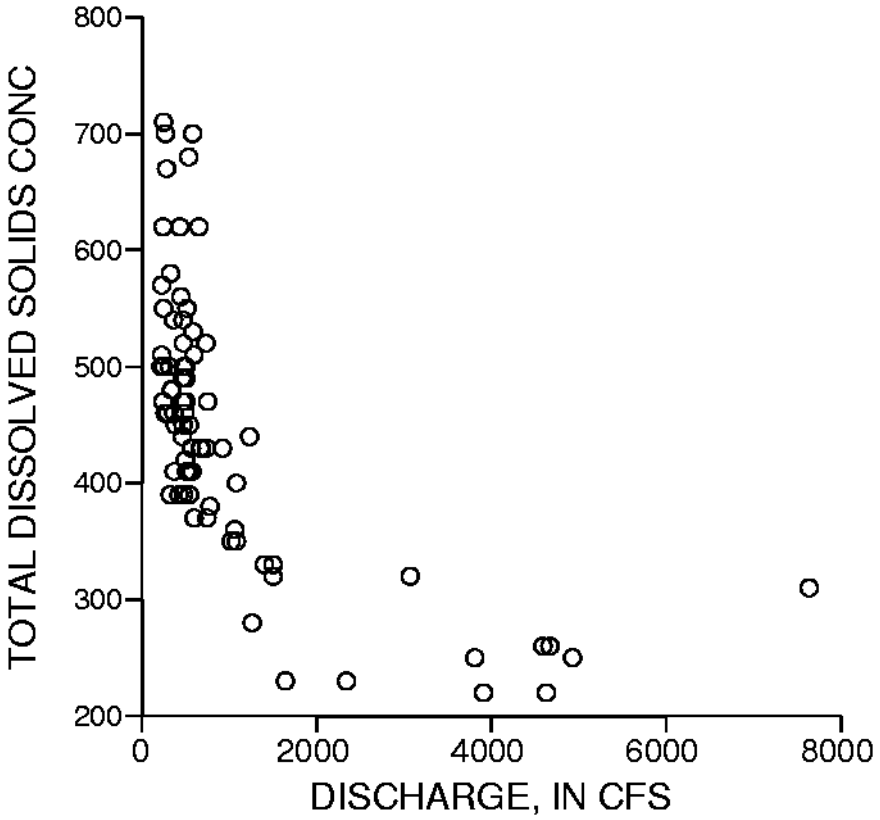


Figure 9.4 Scatterplot of the Cuyahoga R. TDS data

If the problem is curvature only (1a), then try to identify a new x which is a better linear predictor (a transform of the original x or another variable altogether). When possible, use the

best physically-based argument in choosing the right x. It may be appropriate to resort to empirically selecting the x which works best (highest R^2) from among a set of reasonable explanatory variables.

If the problem is non-constant variance, (also called heteroscedasticity, 1b above) or both curvature and heteroscedasticity, then transforming y, or x and y, may be called for. Mosteller and Tukey (1977) provided a guide to selecting power transformations using plots of y versus x called the "bulging rule". Going "up" the ladder of powers means $\theta > 1$ (x^2, etc.) and "down" the ladder of powers means $\theta < 1$ (log x, $1/x$, $\sqrt{x}$, etc.).

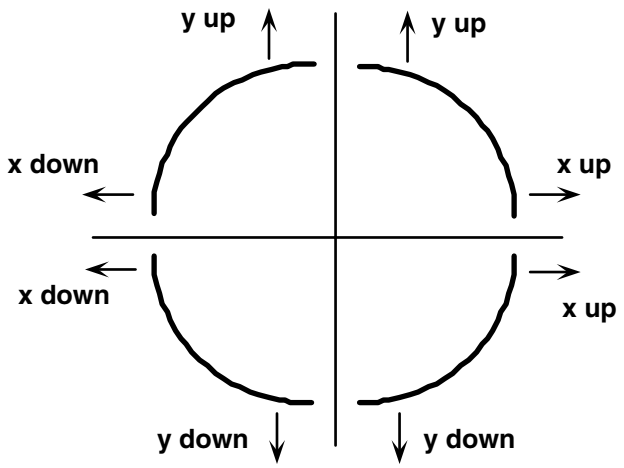


Figure 9.5 The bulging rule for transforming curvature to linearity.
(after Mosteller and Tukey, 1977).

The non-linearity of the TDS data is obvious from figure 9.4, and some type of transformation of the x variable (discharge, denoted Q) is necessary. The base 10 log of Q is chosen, as the plot has the shape of the lower left quadrant of the bulging rule, and so $\theta < 1$. Figure 9.6 presents the TDS data versus the log of Q. Linearity is achieved. There is some hint of greater variance around the line at the lower Q's, but notice that there are also far more data at lower discharges. The range of values can be expected to be greater where there is more data, so non-constant variance is not proven. Therefore this transformation appears acceptable based on the first set of plots.

2) Having selected an appropriate x and y, compute the least squares regression statistics, saving the values of the residuals for further examination. In the regression results, focus on these things:

2a) The coefficients, b_0 and b_1: Are they reasonable in sign and magnitude? Do they lead to predictions of unreasonable values of y for reasonable values of x (e.g., negative flows or concentrations)?

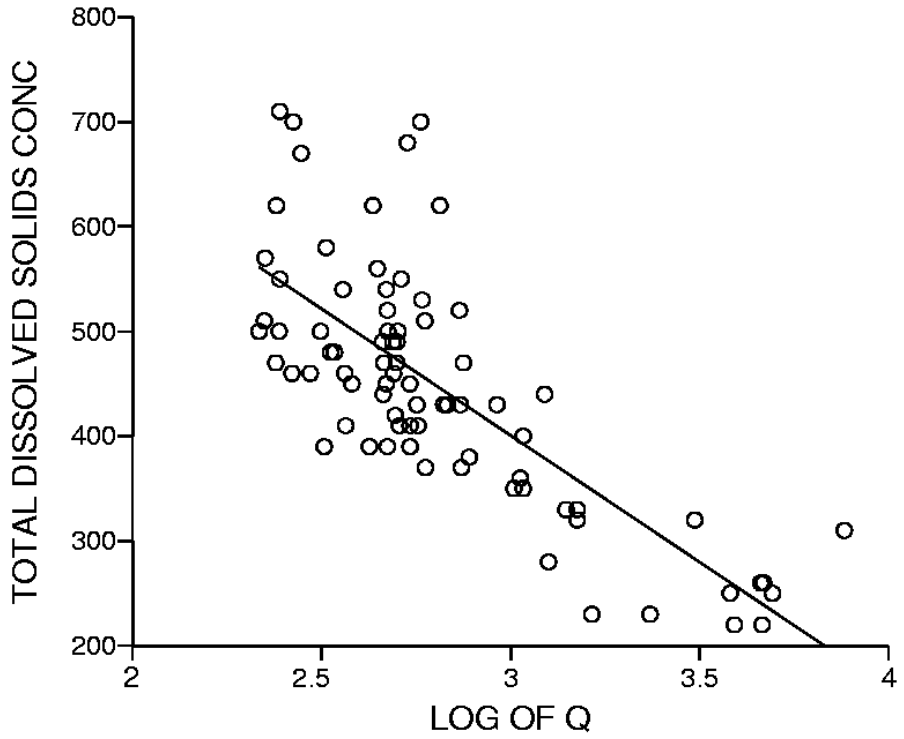


Figure 9.6 Scatterplot with regression line after transformation of x

The Cuyahoga TDS data have the following regression results:

TDS = 1125 – 242 $\log_{10}$ Q

n = 80 s = 75.55 R^2 = 0.57 SS_X = 10.23

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	1125.5	66.9	16.8	0.000
Slope β_1	–241.6	23.6	–10.2	0.000

Table 9.2 Regression statistics for the Cuyahoga TDS data

It appears reasonable that TDS concentrations should be diluted with increasing stream discharge, producing a negative slope. No negative concentrations result from reasonable values for Q at this site.

2b) The R^2: Does the regression explain much variance? Is the amount of variance explained substantial enough to make it worthwhile to use the regression, given the risk that the form of the model is likely to be imperfect? There is no general rule for what is too low an R^2 for a useful regression equation.

For the Cuyahoga data, 57% of the variance of total dissolved solids is explained by the effect of log Q.

2c) Look at the t-ratio (or t-statistics) on the two coefficients. These are the test statistics needed for testing the null hypothesis that the coefficient is equal to zero. In particular, look at the t-ratio on β_1. If $|t|>2$, reject $\beta_1=0$ at $\alpha=0.05$ for reasonably large sample sizes and therefore assert there is a statistically significant linear relationship between x and y. If the t-ratio is between −2 and +2, the observed relationship is no stronger than what is likely to arise by chance alone in the absence of any real linear relationship. If this is the case one should go back to step 1 or give up on the use of regression with this data set. (The formalities of these hypothesis tests are given in a later section).
Both the intercept and slope of the TDS regression are significant at any reasonable α, as shown by the large t-statistics and small p-values of table 9.2.

3) Examine adherence to the assumptions of regression using residuals plots. Three types of residuals plots will clearly present whether or not the regression model adheres sufficiently to the assumptions to be used.

3a) Residuals versus predicted (e vs. $\hat{y}$). Look for two possible problems: curvature and heteroscedasticity. These are exactly the same problems described in step 1. However, plotting residuals enhances the opportunity to see these problems as compared to plotting the original data. The solutions to the problems are the same. Figure 9.7 presents an example of a good residuals plot, one where the residuals show no curvature or changing variance. Figure 9.8, on the other hand, is a residuals plot which shows both curvature and changing variance, producing the typical "horn" pattern which is often correctable by taking the logarithms of y.

It is possible to read too much into these plots, however. Beware of "curvature" produced by a couple of odd points or of error variance seeming to both grow and shrink one or more times over the range of $\hat{y}$. Probably neither of these can or should be fixed by transformation but may indicate the need for the robust procedures of Chapter 10.

In figure 9.9, the residuals from the Cuyahoga TDS regression are plotted versus its predicted values. There is an indication of heteroscedasticity, though again there are more data for the larger predicted values. There also appears to be a bow in the data, from + to − and back to + residuals. Perhaps a transformation of the TDS concentrations are warranted, or the incorporation of additional variables into the regression equation.

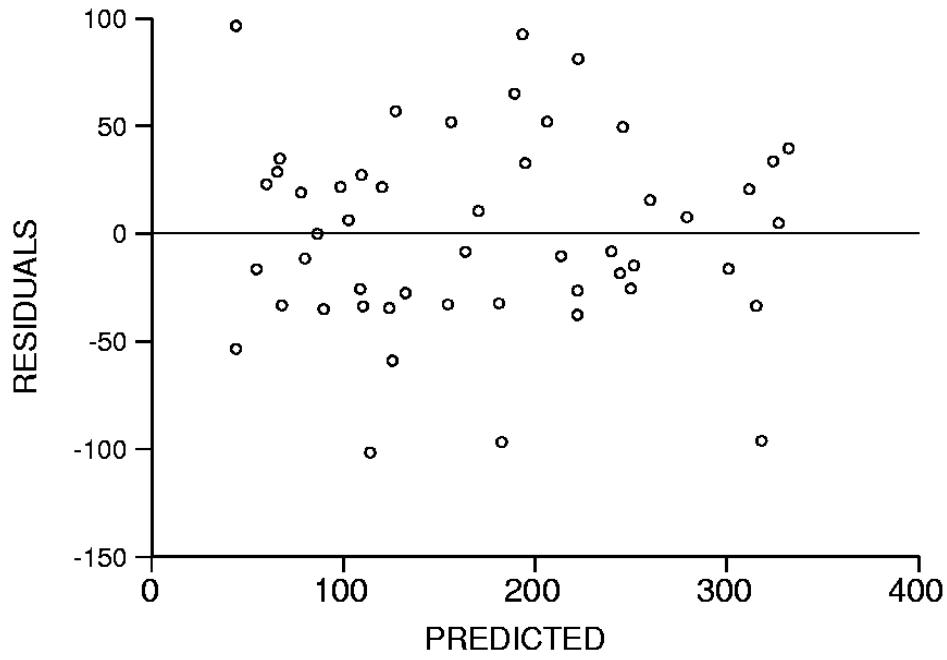


Figure 9.7 Example of a residuals plot for a good regression model

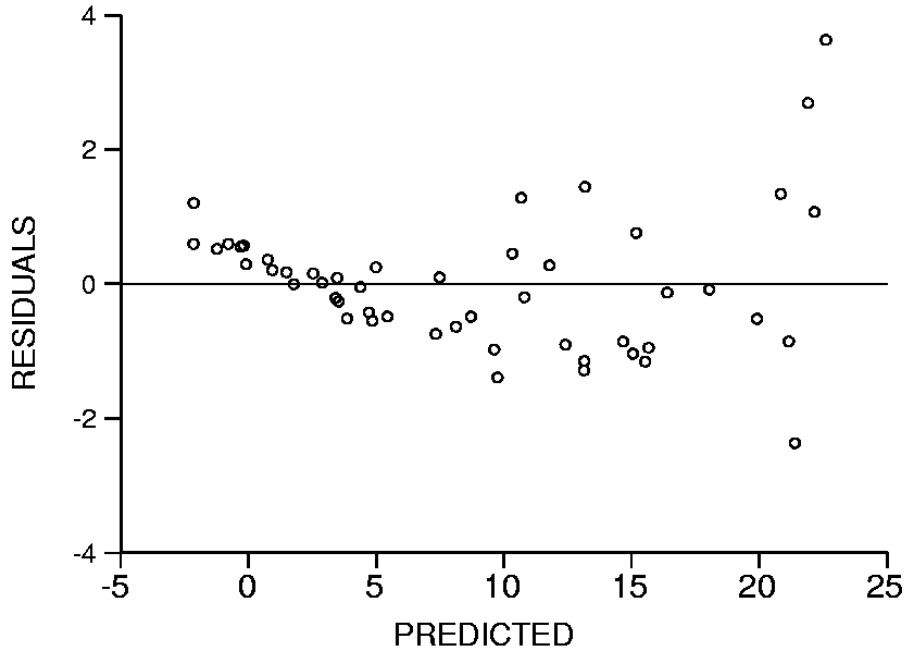


Figure 9.8 Residuals plot showing curvature and changing variance.

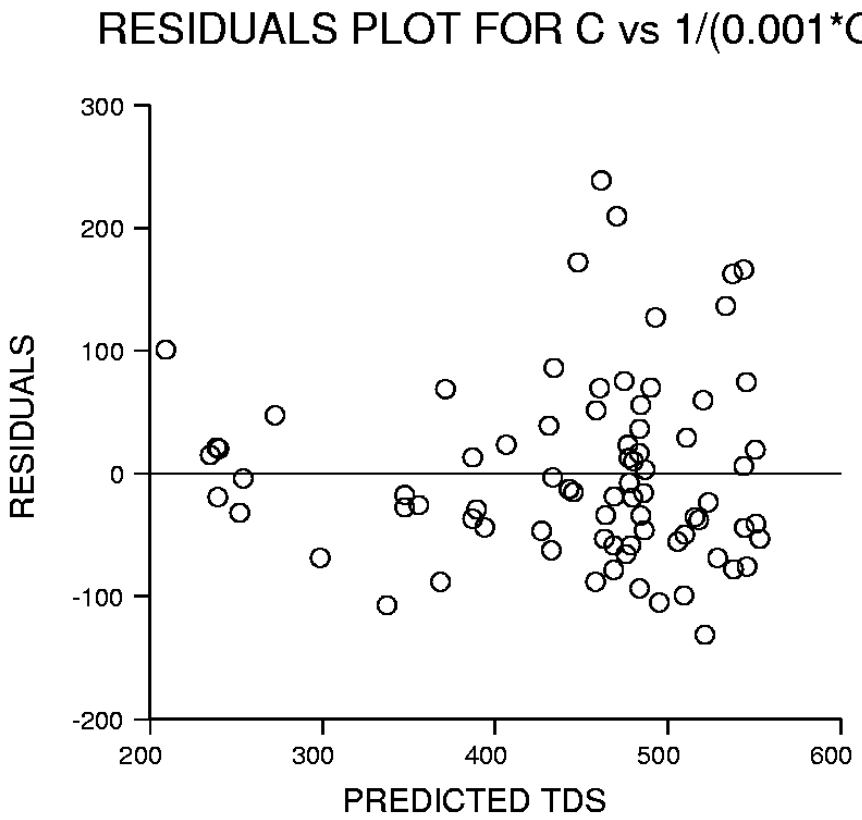


Figure 9.9 Residuals plot of the Cuyahoga data.

3b) Residuals versus time (e vs. t). If there is any time or space order to the observations (relating to time of collection, time of measurement, or map location), plot the residuals versus time or season or time of day, or versus the appropriate 1- or 2-dimensional space coordinate to see if there is a pattern in the residuals. A good residuals pattern, one with no relation between residuals and time, will look similar to figure 9.7 -- random noise. If on the other hand structure in the pattern over time is evident, seasonality, long-term trend, or correlation in the residuals may be the cause. Trend or seasonality suggest adding a new term to the regression equation (see Chapter 12). Correlation between residuals over time or space require one of the remedies listed in section 9.5.4.

Correlation between residuals over time or space may not be evident from the e_i versus $\hat{y}$ residuals plot (figure 9.10a), but will stand out on a plot of e_i versus time (9.10b). The nonrandomness is evident in that positive residuals clump together, as do negative -- a positive correlation. Plotting the ith versus the (i–1)th residual shows this pattern more strongly (9.10c). If time or space are measured as categorical variables (month, etc.), plot boxplots of residuals by category and look for patterns of regularity. Where no differences occur between boxes, the time or space variable has no effect on the response variable.

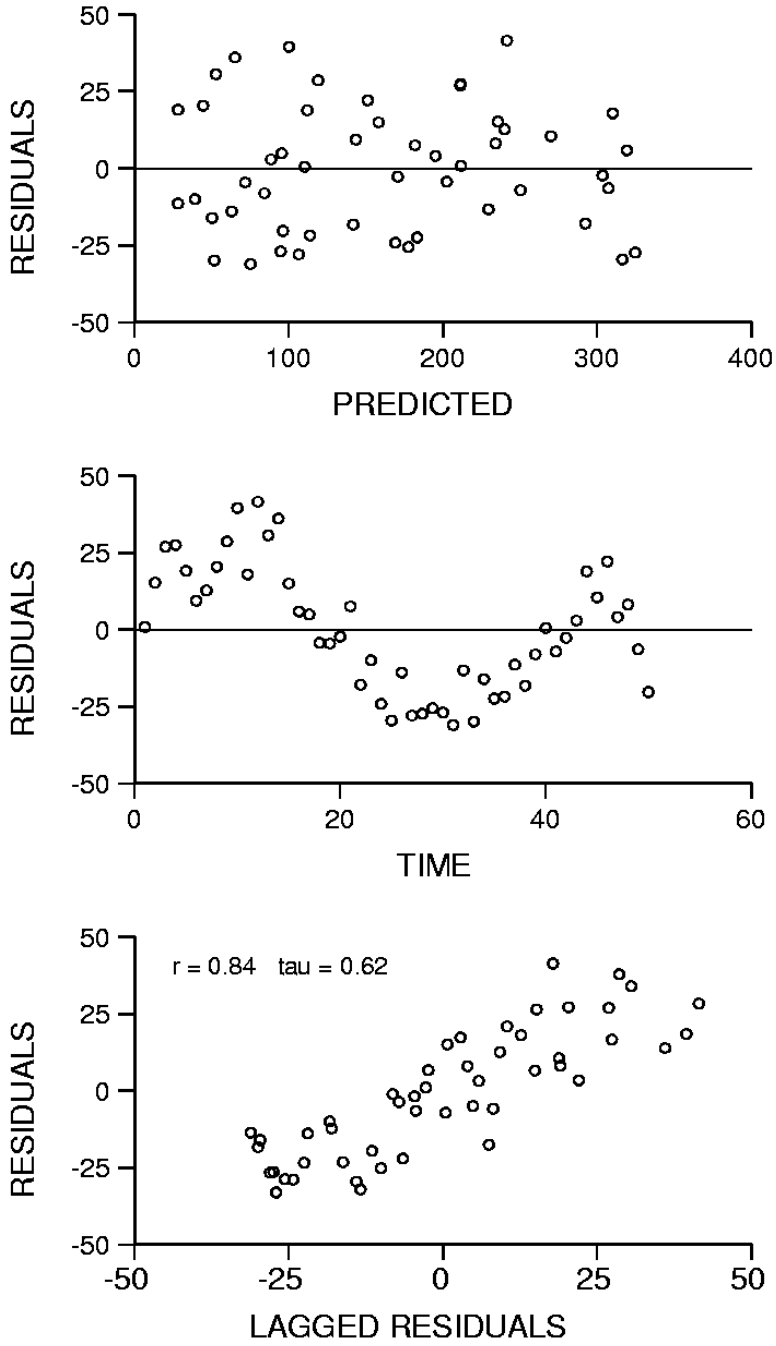


Figure 9.10 a) Residual e_i vs $\hat{y}$ plot shows no hint of correlation over time
b) Time series of residuals shows e_i related to time
c) Correlation of e_i vs. $e_{(i-1)}$

In figure 9.11, boxplots of TDS residuals by month show a definite seasonality, with generally high residuals occurring in the winter months, low residuals in the summer, and unusually high values in September. Thus the regression equation will underpredict concentrations in the

winter and overpredict in the summer. This pattern may be due to washoff of road de-icing salts in the winter. The unknown cause of the September anomalies should be investigated further. To better mimic the seasonal variation, other explanatory variables must be added. This will be discussed in Chapter 12.

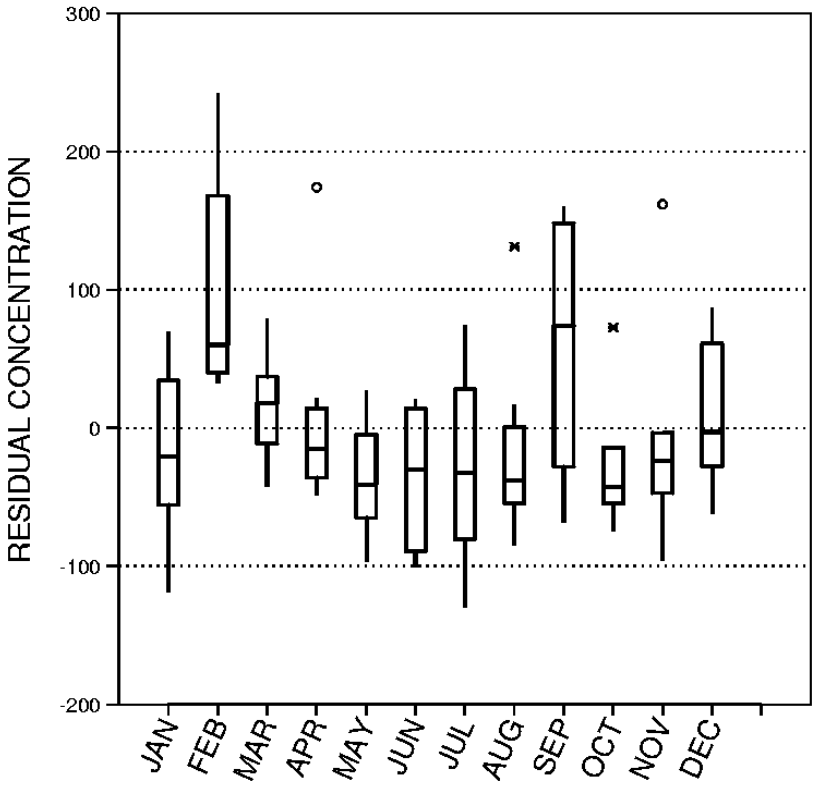


Figure 9.11 Residual of TDS concentrations by month. Note the seasonality.

3c) Normality of residuals. Examine the distribution of residuals using a boxplot, stem and leaf, histogram, or normal probability plot. If they depart very much from a normal distribution, then the various confidence intervals, prediction intervals, and tests described below will be inappropriate. Specifically,
(i) hypothesis tests will have low power (slopes or explanatory variables will falsely be declared insignificant), and
(ii) confidence or prediction intervals will be too wide, as well as giving a false impression of symmetry.

A boxplot of residuals from the TDS-logQ regression shown in figure 9.12 is mildly right-skewed, with several outliers present. A probability plot of the residuals (figure 9.13) shows a slight departure from normality. If these were the only problems, transformation of the y variable might not be warranted. But combined with the

problems already noted above of curvature and heteroscedasticity, further work is required.

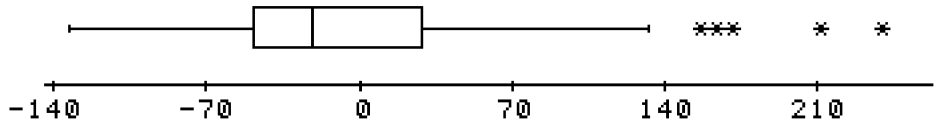


Figure 9.12 Boxplot of the TDS regression residuals

For further attempts to find an appropriate transformation of the Cuyahoga data, see problem 9.1 at the end of this chapter.

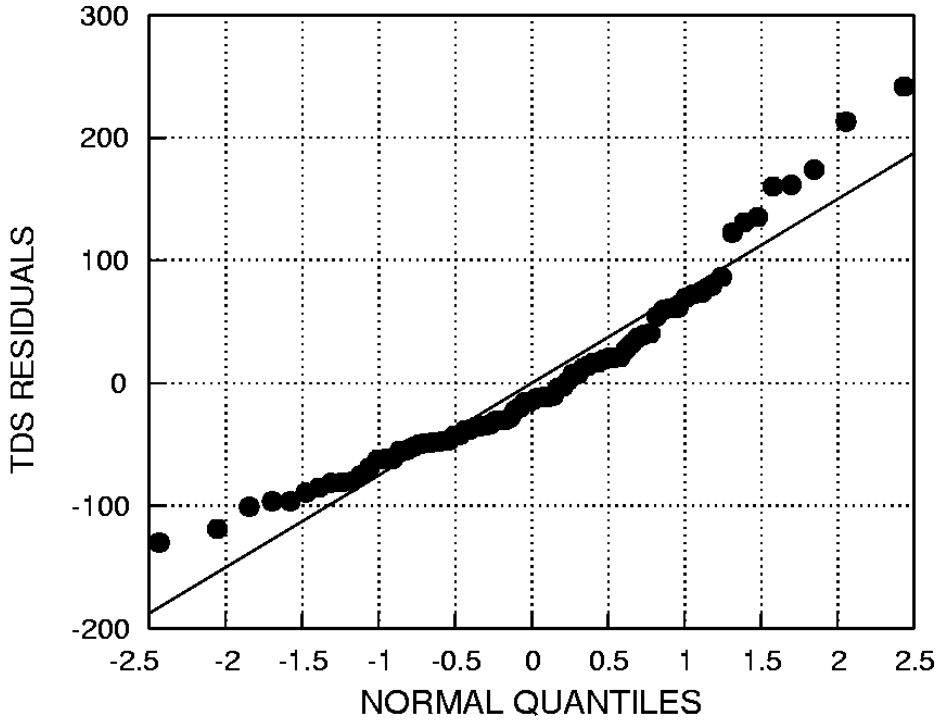


Figure 9.13 Probability plot of the TDS regression residuals

3d) Residuals versus other explanatory variables. To determine whether other explanatory variables should be included into a multiple regression model, boxplots of residuals by categorical explanatory variables or scatterplots versus continuous variables should be plotted. If something other than a random pattern occurs, that variable or one like it may be appropriate for adding to the regression equation. Figure 9.14 for example might result from plotting residuals from a regression of radon concentrations in water versus uranium content of rocks. using different symbols for wells and springs. The residuals for wells tend to be larger than those for springs, as also shown by the boxplots at the side. Incorporating an additional explanatory variable for "water source" into the

regression equation using the techniques of Chapter 11 explains more of the noise in the data, improving the model.

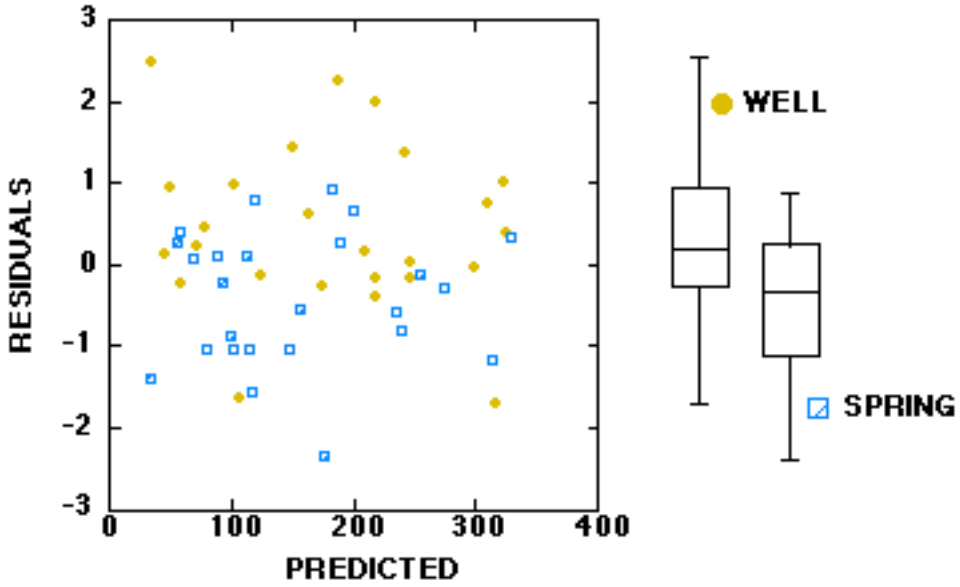


Figure 9.14 Residuals plotted by an additional explanatory variable.

4) Use the regression diagnostics of section 9.5 to ensure that one or two observations are not strongly influencing the values of the coefficients, and to determine the quality of predicted values. These diagnostics duplicate much of what can be seen with plots for a single explanatory variable, but become much more important when performing multiple regression.

9.4 Hypothesis Testing in Regression

9.4.1 Test for Whether the Slope Differs From Zero

The hypothesis test of greatest interest in regression is the test for a significant slope (β_1). Typically, the null hypothesis is

H_0: $\beta_1 = 0$

versus the alternative hypothesis

H_1: $\beta_1 \neq 0$.

The null hypothesis also states that the value of y does not vary as a linear function of x. Thus for the case of a single explanatory variable this also tests for whether the regression model has statistical significance. A third interpretation is as a test for whether the linear correlation coefficient significantly differs from zero. The latter two interpretations are <u>not</u> applicable for

multiple explanatory variables. The test statistic computed is the t-ratio (the fitted coefficient divided by its standard error):

$$t = \frac{b_1}{s/\sqrt{SS_x}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

H_0 is rejected if $|t| > t_{crit}$, where t_{crit} is the point on the Student's t distribution with n–2 degrees of freedom, and with probability of exceedance of $\alpha/2$.

Note that when $\alpha=0.05$ and $n>30$ $t_{crit} \cong 2.0$

and when $\alpha=0.01$ and $n>30$ $t_{crit} \cong 2.6$.

For the Cuyahoga TDS example the t-statistic for β_1 was much greater than 2, and indeed was significant at the $\alpha = 0.0001$ level. Therefore a strong linear correlation exists between TDS and $\log_{10}$ of Q.

This test for nonzero slope can also be generalized to testing the null hypothesis that $\beta_1=\beta_1^*$ where β_1^* is some pre-specified value. For this test the statistic is defined as

$$t = \frac{b_1 - b_1^*}{s/\sqrt{SS_x}}$$

9.4.2 Test for Whether the Intercept Differs from Zero

Tests on the intercept b_0 can also be computed. The test for

$H_0: b_0 = 0$

is usually the one of interest. The test statistic is

$$t = \frac{b_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}}$$

H_0 is rejected if $|t| > t_{crit}$ where t_{crit} is defined as in the previous test. From table 9.2 the intercept for the TDS data is seen to be highly significantly different from 0.

It can be dangerous to delete the intercept term from a regression model. Even when the intercept is not significantly different from zero, there is little benefit to forcing it to equal zero, and potentially great harm in doing so. Regression statistics such as R^2 and the t-ratio for β_1 lose their usual meaning when the intercept term is dropped (set equal to zero). Recognition of a physical reason why y must be zero when x is zero is not a sufficient argument for setting $b_0 = 0$. Probably the only appropriate situation for fitting a no-intercept model is when all of the following conditions are met:

1) the x data cover several orders of magnitude,
2) the relationship clearly looks linear from zero to the most extreme x values,
3) the null hypothesis that $\beta_0 = 0$ is not rejected, and
4) there is some economic or scientific benefit to dropping the intercept.

9.4.3 Confidence Intervals on Parameters

Confidence intervals for the individual parameters β_0, β_1, and σ^2 indicate how well these can be estimated. The meaning of the $(1-\alpha)\bullet 100\%$ confidence interval is that, in repeated collection of new data and subsequent regressions, the frequency with which the true parameter value would fall outside the confidence interval is α. For example, $\alpha = 0.05$ confidence intervals around the estimated slopes of the regression lines in figure 9.3 would include the true slope 95% of the time.

For the slope β_1 the confidence interval (C.I.) is

$$\left(b_1 - \frac{t\, s}{\sqrt{SS_x}} \;,\; b_1 + \frac{t\, s}{\sqrt{SS_x}} \right)$$

where t is the point on the student's t-distribution having $n-2$ degrees of freedom with a probability of exceedance of $\alpha/2$.

For the intercept β_0 the C.I. is

$$\left(b_0 - ts\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}, b_0 + ts\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}} \right)$$

where t is defined as above.

For the variance σ^2 (also called the mean square error MSE), the C.I. is

$$\left(\frac{(n-2)s^2}{\chi^2_{1-a/2}}, \frac{(n-2)s^2}{\chi^2_{a/2}} \right)$$

where χ^2_p is the quantile of the chi-square distribution having $n-2$ degrees of freedom with exceedance probability of p.

As an example, the 95% confidence intervals for the Cuyahoga TDS data are:

For β_1: $\left(-241.6 - \frac{1.99\bullet 75.6}{\sqrt{10.23}} \;,\; -241.6 + \frac{1.99\bullet 75.6}{\sqrt{10.23}} \right) = (-288.6 \,,\, -194.6)$

For β_0: $\left(1125.5 - 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{2.81^2}{10.23}} \;,\; 1125.5 + 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{2.81^2}{10.23}} \right)$

$= (991.8 \,,\, 1258.7)$

For σ^2: $\left(\frac{(78)\ 5708}{104.3} , \frac{(78)\ 5708}{55.5} \right) = (4269 , 8022)$

9.4.4 Confidence Intervals for the Mean Response

There is also a confidence interval for the conditional mean of y given any value of x. If x_0 is a specified value of x, then the estimate of the expected value of y at x_0 is $\hat{y} = b_0 + b_1 x_0$, the value predicted from the regression equation. But there is some uncertainty to this, associated with the uncertainty for the true parameters β_0 and β_1. The $(1-\alpha)\bullet 100\%$ confidence interval for the mean y is then

$$\left(\hat{y} - ts\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}},\ \hat{y} + ts\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}} \right)$$

where t is the quantile of the students' t-distribution having n−2 degrees of freedom with probability of exceedance of $\alpha/2$. Note that the confidence interval is two-sided, requiring a t-statistic of $\alpha/2$ for either side. Also note from the formula that the farther x_0 is from $\bar{x}$ the wider the interval becomes. That is, the model is always "better" near the middle of the x values than at the extremes.

To continue the Cuyahoga TDS example, the confidence interval for the mean y is calculated for two values of x_0, 3.0 (near $\bar{x}$) and 3.8 (far from $\bar{x}$):

for $x_0 = 3.0$: $\left(399 - 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{(3.0\text{-}2.81)^2}{10.23}} , 399 + 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{(3.0\text{-}2.81)^2}{10.23}} \right)$

$= (380 , 418)$

for $x_0 = 3.8$: $\left(205.4 - 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{(3.8\text{-}2.81)^2}{10.23}} , 205.4 + 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{(3.8\text{-}2.81)^2}{10.23}} \right)$

$= (155.9, 254.9)$

a confidence interval of width 38 at $x_0 = 3.0$, and a width of 99 at $x_0 = 3.8$.

When the confidence interval for each logQ value is connected together, the characteristic "bow" shape of regression confidence intervals can be seen (figure 9.15). Note that this shape agrees with the pattern seen in figure 9.3 for randomly generated regression lines, where the positions of the line estimates are more tightly controlled near the center than near the ends.

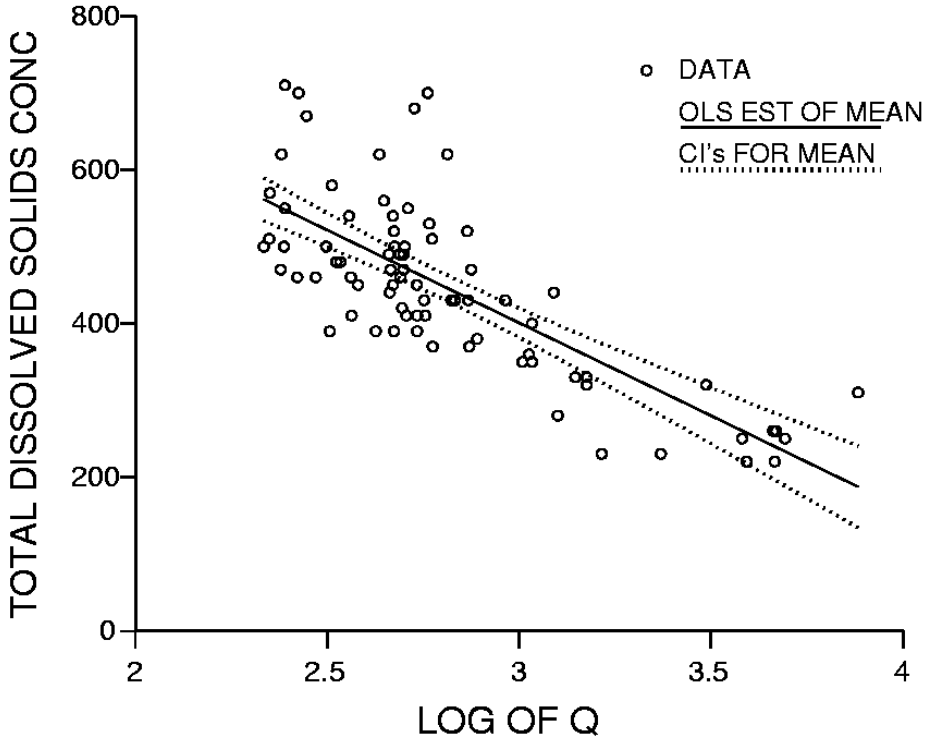


Figure 9.15 Confidence intervals for mean TDS for the Cuyahoga River data.

9.4.5 Prediction Intervals for Individual Estimates of y

The prediction interval, the confidence interval for prediction of an estimate of an individual y, is often confused with the confidence interval for the mean. This is not surprising, as the best estimate for both the mean of y given x_0 and for an individual y given x_0 are the same -- $\hat{y}$. However, their confidence intervals differ. The formulas are identical except for one very important term. The prediction interval incorporates the unexplained variability of y (σ^2) in addition to uncertainties in the parameter estimates β_1 and β_2. The $(1-\alpha)\bullet 100\%$ prediction interval for a single response is

$$\left(\hat{y} - ts\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}}, \quad \hat{y} + ts\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}} \right)$$

where all of the terms are as defined previously. Note that these intervals widen as x_0 departs from x, but not nearly as markedly as the confidence intervals do. In fact, a simple rough approximation to the prediction interval is just ($\hat{y}$ − ts, $\hat{y}$ + ts), two parallel straight lines. This is because the second and third terms inside the square root are negligible in comparison to the first, provided the sample size is large. These prediction intervals should contain approximately $1-\alpha\bullet(100)\%$ of the data within them, with $\alpha/2\bullet(100)\%$ of the data beyond each side of the intervals. They will do so if the residuals are approximately normal.

The prediction intervals for the Cuyahoga TDS data are plotted in figure 9.16. They are computed below for $x_0 = 3.0$ and 3.8.

for $x_0 = 3.0$:

$$\left(399 - 1.99 \bullet 75.6\sqrt{1+\frac{1}{80}+\frac{(3.0-2.81)^2}{10.23}}\ ,\ 399 + 1.99 \bullet 75.6\sqrt{1+\frac{1}{80}+\frac{(3.0-2.81)^2}{10.23}}\right)$$
$$= (247.4\ ,\ 550.6)$$

for $x_0 = 3.8$:

$$\left(205.4 - 1.99 \bullet 75.6\sqrt{1+\frac{1}{80}+\frac{(3.8-2.81)^2}{10.23}}\ ,\ 205.4 + 1.99 \bullet 75.6\sqrt{1+\frac{1}{80}+\frac{(3.8-2.81)^2}{10.23}}\right)$$
$$= (47.0,\ 363.8)$$

a prediction interval of width = 303 at $x_0 = 3.0$, and a width of 317 at $x_0 = 3.8$. Note that the prediction intervals are much wider than the confidence intervals, and that there is only a small difference in width between the two prediction intervals as x_0 changes. Also note from figure 9.16 that the data appear skewed, with all of the values found beyond the prediction intervals falling above the upper interval.

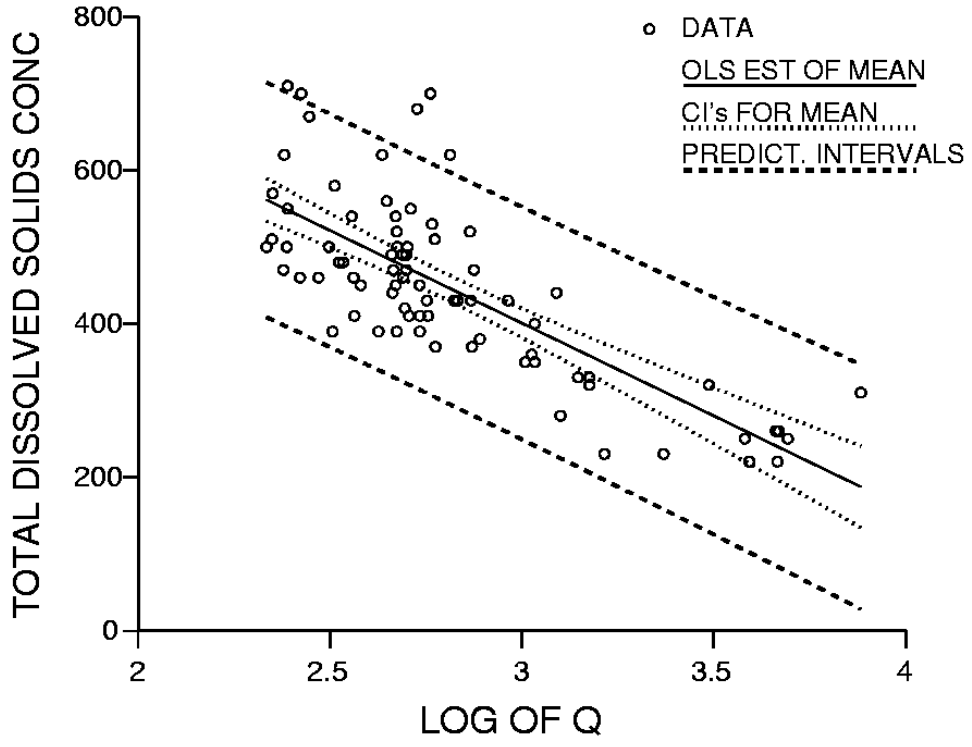


Figure 9.16 Prediction intervals for an individual TDS estimate -- Cuyahoga River.

9.4.5.1 Nonparametric prediction interval

There is also a nonparametric version of the prediction interval. This might be used when the x,y data display a linear relationship and residuals have constant variance (homoscedastic), but the distribution of the residuals appears non-normal. Typically, such departures from normality take the form of skewness or an excessive number of outside or far outside values (as seen in a boxplot). The nonparametric prediction interval is

$$(\hat{y} + e_{(L)},\ \hat{y} + e_{(U)})$$

where $e_{(L)}$ and $e_{(U)}$ are the $1-\alpha/2$ and $\alpha/2$th quantiles of the residuals.

In other words, $e_{(L)}$ is the Lth ranked residual and $e_{(U)}$ is the Uth ranked residual, where $L = (n+1)\bullet\alpha/2$ and $U = (n+1)\bullet(1-\alpha/2)$. When L and U are not integers either the integer closest to L and U can be chosen, or $e_{(L)}$ and $e_{(U)}$ can be interpolated between adjacent residuals.

For the Cuyahoga TDS data, $L = 81\bullet.025 = 2.025$ and $U = 81\bullet.975 = 78.975$. Either the 2nd and 79th ranked residual can be selected, or values interpolated between the 2nd and 3rd, and the 78th and 79th residual. These are then added to the regression line ($\hat{y}$). In figure 9.17 the nonparametric prediction interval is compared to the one previously developed assuming normality of residuals. Note that the nonparametric interval is asymmetric around the central regression line, reflecting the asymmetry of the data.

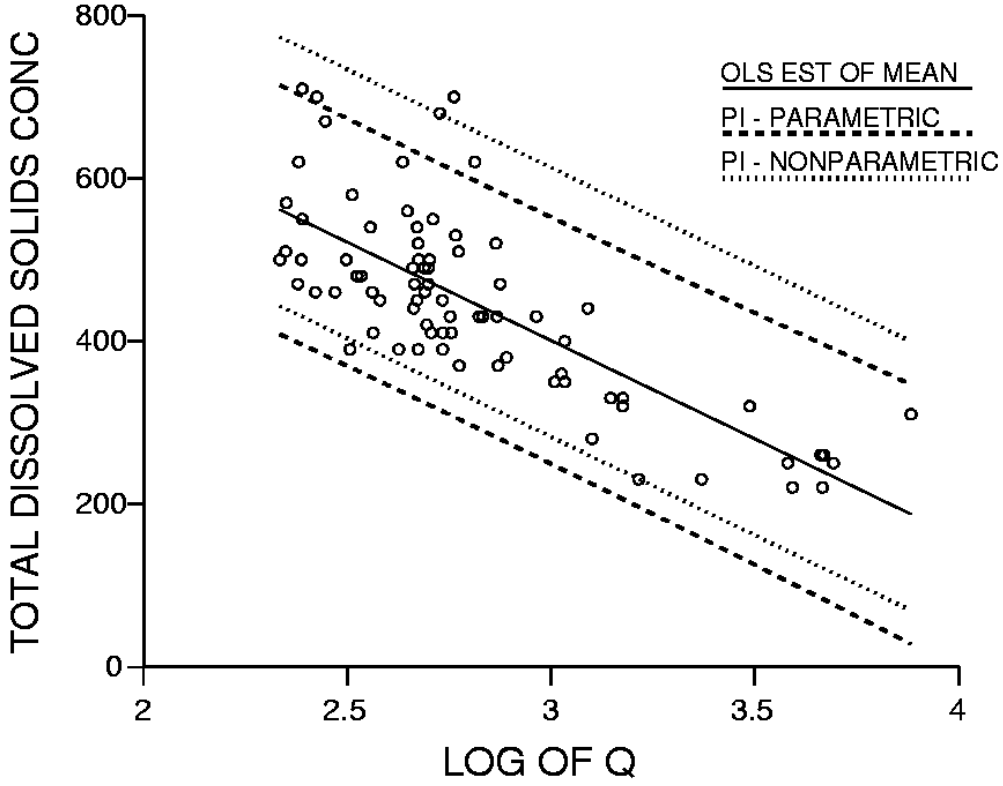


Figure 9.17 Nonparametric and parametric prediction intervals for the TDS data.

9.5 Regression Diagnostics

One common mistake in regression analysis is to base decisions about model adequacy solely on the regression summary statistics--principally R^2, s and the F- or t-test results. R^2 is a measure of the percent of the variation in the response (y) variable that is accounted for by the variation in the explanatory variables. The s (standard error of the regression or standard deviation of the residuals) is a measure of the dispersion of the data around the regression line. Most regression programs also perform an overall F-test to determine if the regression relationship is statistically significant, ie. that the apparent relationship between y and x is not likely to arise due to chance alone. Some programs also do a t-test for each explanatory variable to determine if the coefficient for that variable is significantly different from zero.

These statistics provide substantial information about regression results. An equation that accounts for a large amount of the variation in the response variable and has coefficients that are statistically significant is highly desirable. However, decisions about model adequacy cannot be made on the basis of these criteria alone. A large R^2 or significant F-statistic does not guarantee that the data have been fitted well. Figure 9.18 (Anscombe, 1973) illustrates this point.

The data in the four graphs have exactly the same summary statistics and regression line (same b_0, b_1, s, R^2). In 9.18a is a perfectly reasonable regression model, an evidently linear relationship having an even distribution of data around the least-squares line. The strong curvature in 9.18b suggests that a linear model is highly inadequate and that some transformation of x would be a better explanatory variable, or that an additional explanatory variable is required. With these improvements perhaps all of the variance could be explained. Figure 9.18c illustrates the effect of a single outlier on regression. The line mis-fits the data, and is drawn towards the outlier. Such an outlier must be recognized and carefully examined to verify its accuracy if possible. If it is impossible to demonstrate that the point is erroneous, a more robust procedure than regression should be utilized (see Chapter 10). The regression slope in 9.18d is strongly affected by a single point (the high x value), with the regression simply connecting two "points", a single point plus a small cluster of points. Such situations often produce R^2 values close to 1, yet may have little if any predictive power. Had the outlying point been in a different location, the resulting slope would be totally different. For example, the only difference between the data of figure 9.19a and 9.19b is the rightmost data point. Yet the slopes are entirely different! Regression should not be used in this case because there is no possible way to evaluate the assumptions of linearity or homoscedasticity without collecting more data in the gap between the point and cluster. In addition, the slope and R^2 are totally controlled by the position of one point, an unstable situation.

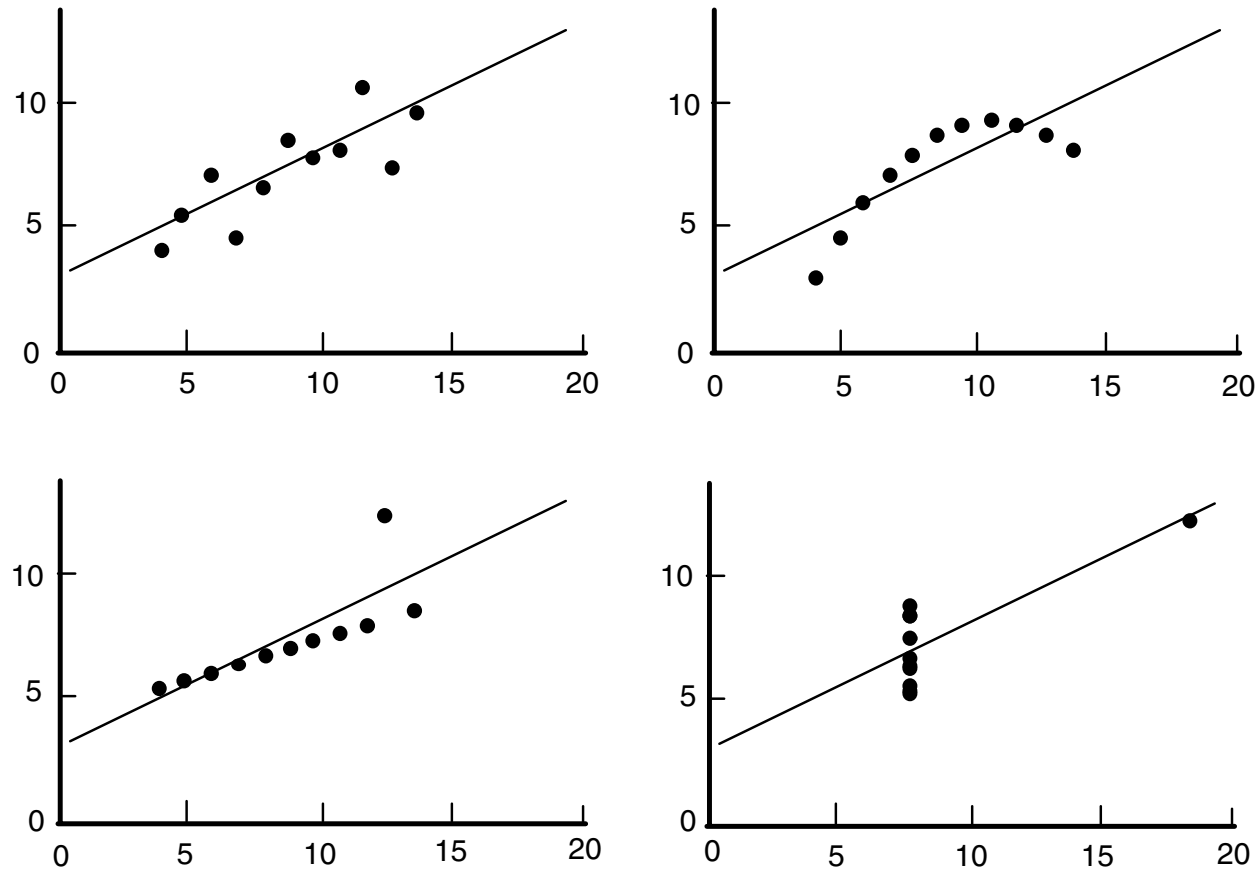


Figure 9.18 Three key pathologies in regression (after Anscombe, 1973).

The three key pathologies can be referred to by simple names: curvature (9.18b), outlier or large residual (9.18c), and high influence and leverage (9.18d). They are generally easy to identify from plots (y vs. x, or e vs. $\hat{y}$) in a linear regression with one explanatory variable. However, in multiple linear regression they are much more difficult to visualize or identify, requiring plots in multi-dimensional space. Thus numerical measures of their occurrence, called "regression diagnostics", have been developed.

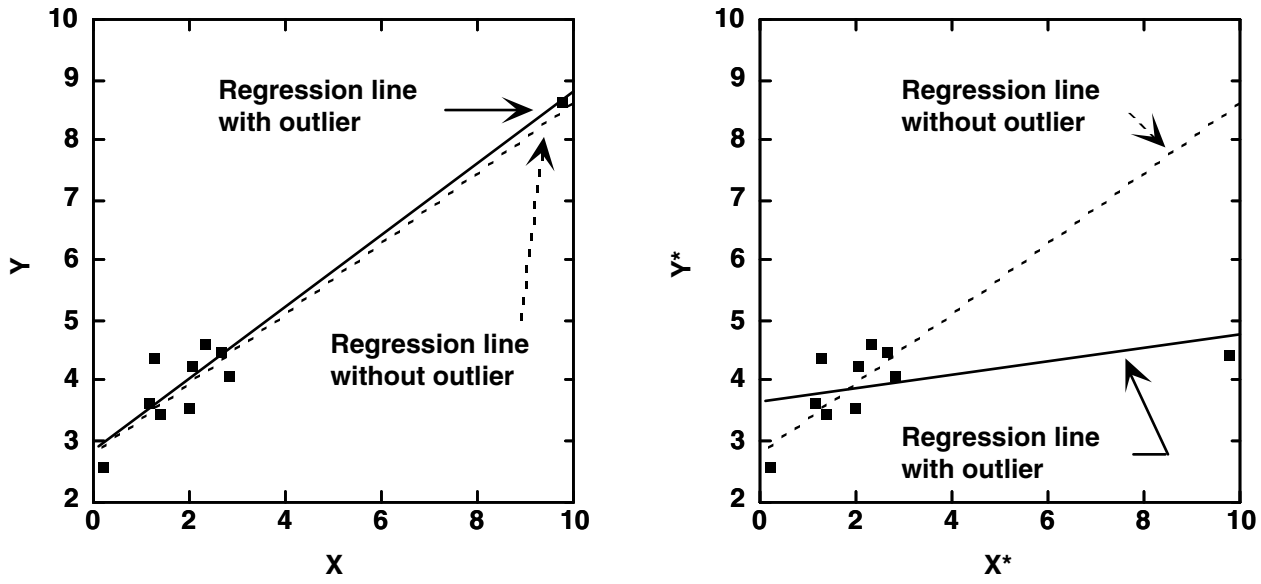
Equations for diagnostics useful in identifying points of leverage, influence, or outliers are given here in terms of the two dimensions (x,y) applicable to simple linear regression (SLR). Each can be generalized using matrix notation to a larger number of dimensions for multiple linear regression (MLR). Further references on regression diagnostics are Belsley, Kuh, and Welsch (1980), Draper and Smith (1981), and Montgomery and Peck (1982).

9.5.1 Measures of Outliers in the x Direction

9.5.1.1 Leverage

Leverage is a measure of an "outlier" in the x direction, as in graph 9.18a. It is a function of the distance from the ith x value to the middle (mean) of the x values used in the regression. Leverage is usually denoted as h_i, the ith diagonal term of the "hat" matrix $X (X'X)^{-1} X'$, or for SLR

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_x}.$$



a) outlier with high leverage, but low influence

b) outlier with high leverage and high influence

Figure 9.19 Influence of location of a single point on the regression slope.

A high leverage point is one where $h_i > 3p/n$ where p is the number of coefficients in the model (p=2 in SLR, b_0 and b_1). Though leverage is concerned only with the x direction, a high leverage point has the potential for exerting a strong influence on the regression slope. If the high leverage point falls far from the regression line that would be predicted if it were absent from the data set, then it is a point with high influence as well as high leverage (figure 9.19b).

9.5.2 Measures of Outliers in the y Direction

9.5.2.1 Standardized residual

One measure of outliers in the y direction is the standardized residual e_{si}. It is the actual residual $e_i = y_i - \hat{y}_i$ standardized by its standard error.

$$e_{si} = \frac{e_i}{s\sqrt{1 - h_i}}$$

An extreme outlier is one for which $|e_{si}|>3$. There should be only an average of 3 of these in 1,000 observations if the residuals are normally distributed. $|e_{si}|>2$ should occur about 5 times in 100 observations if normally distributed. More than this number indicates that the residuals do not have a normal distribution.

9.5.2.2 Prediction residuals and the PRESS statistic

A very useful form of residual computation is the prediction residual $e_{(i)}$. These are computed as $e_{(i)} = y_i - \hat{y}_{(i)}$ where $\hat{y}_{(i)}$ is the regression estimate of y_i based on a regression equation computed leaving out the ith observation. The (i) symbolizes that the ith observation is left out of the computation. These are easily calculated using leverage statistics without having to perform n separate regressions:

$$e_{(i)} = e_i / (1 - h_i) .$$

One of the best measures of the quality of a regression equation is the "PRESS" statistic, the "PRediction Error Sum of Squares."

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2$$

PRESS is a validation-type estimator of error. Instead of splitting the data set in half, one-half to develop the equation and the second to validate it, PRESS uses n−1 observations to develop the equation, then estimates the value of the one left out. It then changes the observation left out, and repeats the process for each observation. The prediction errors are squared and summed. Minimizing PRESS means that the equation produces the least error when making new predictions. In multiple regression it is a very useful estimate of the quality of possible regression models.

9.5.2.3 Studentized residuals

Studentized residuals (TRESIDs) are used as an alternate measure of outliers by some texts and computer software. They are often confused with standardized residuals.

$$\text{TRESID}_i = \frac{e_i}{s_{(i)}\sqrt{1-h_i}} = \frac{e_{(i)}\sqrt{1-h_i}}{s_{(i)}}$$

where

$$s^2_{(i)} = \frac{(n-p)\,s^2 - [\,e_{(i)}^2 / (1 - h_i)\,]}{n - p - 1}$$

TRESIDs are often similar to the standardized residuals e_{si}, but are computed using a variance $s^2_{(i)}$ which does not include their own observation. Therefore an unusually large observation does not inflate the estimate of variance used to determine whether that observation is unusual, and outliers are more easily detected. Under a correct model with normal residuals, TRESIDs have the theoretical advantage that they should follow a t-distribution with n−p−1 degrees of freedom.

9.5.3 Measures of Influence

Observations with high influence are those which have both high leverage and large outliers (figure 9.19b). These exert a stronger influence on the position of the regression line than other observations.

9.5.3.1 Cook's D

One of the most widely used measures of influence is "Cook's D" (Belsley et al., 1980).

$$D_i = \frac{e_i^2 \, h_i}{p s^2 (1 - h_i)^2} = \frac{e_{(i)}^2 \, h_i}{p s^2}$$

The ith observation is considered to have high influence if $D_i > F_{(p+1,n-p)}$ at $\alpha=0.1$ where p is again the number of coefficients. Note that, for SLR with more than about 30 observations, the critical value for D_i would be about 2.4, and for MLR with several explanatory variables the critical value would be in the range of 1.6 to 2.0. Finding an observation with high Cook's D should lead to a very careful examination of the data value for possible errors or special conditions which might have prevailed at the time it occurred. If it can be shown that an error occurred, the point should be corrected if possible, or deleted if the error can't be corrected. If no error can be proven, two options can be considered. A more complex model which fits the point better is one option. The second option is to use a more robust procedure such as that based on Kendall's τ (for one x variable) or weighted least squares (for more than one x variable). These methods for "robust regression" are discussed in Chapter 10.

9.5.3.2 DFFITS

The second influence diagnostic, related to TRESIDs, is the DFFITS:

$$DFFITS_i = \frac{e_i \sqrt{h_i}}{s_{(i)} (1-h_i)} = \frac{e_{(i)} \sqrt{h_i}}{s_{(i)}}$$

An observation is considered to have high influence if $|DFFITS_i| \geq 2\sqrt{p/n}$.

The identification of outliers can be done with either standardized or studentized residuals, and the identification of highly influential points can be done with either DFFITS or Cook's D. The leverage statistic identifies observations unusual in x. PRESS residuals are rarely used except to sum into the PRESS statistic, in order to compare competing multiple regression models.

Example 1

The data of figure 9.19a were analyzed by regression, and the above diagnostics calculated. These data exhibit high leverage but low influence, as removal of the one outlier in the x direction will not appreciably alter the slope of the regression line. The regression results are given in Table 9.3. The only unusual value is the leverage statistic h_i for the last point, the one which plots to the right on the graph. A value of
$3p/n = 0.6$, so the 0.919 for this point shows it to be one of high leverage.

$y = 2.83 + 0.60\ x$

$n = 10$ $s = 0.43$ $R^2 = 0.94$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	2.828	0.195	14.51	0.000
Slope β_1	0.596	0.054	10.98	0.000

OBS#	e_i	h_i	$e_{(i)}$	e std	e stud	DFFITS	D_i
1	−0.377	0.188	−0.465	−0.974	−0.970	−0.467	0.110
2	0.085	0.131	0.098	0.213	0.200	0.077	0.003
3	0.804	0.126	0.920	1.997	2.640	1.005	0.289
4	−0.219	0.122	−0.249	−0.543	−0.518	−0.193	0.020
5	−0.484	0.104	−0.541	−1.189	−1.226	−0.419	0.082
6	0.204	0.104	0.228	0.501	0.476	0.162	0.014
7	0.380	0.101	0.423	0.931	0.922	0.309	0.048
8	0.059	0.100	0.066	0.146	0.136	0.045	0.001
9	−0.462	0.101	−0.514	−1.132	−1.156	−0.388	0.072
10	0.010	**0.919**	0.132	0.087	0.081	0.276	0.043

Table 9.3 Regression statistics for the data of Figure 9.19a

Table 9.4 presents the analysis of the data for figure 9.19b. Note that the equation and ensuing R^2 are quite different. Only y for the 10th observation was changed from its previous value. Note also that the influence statistics DFFITS and D_i are large. The 10th observation is one of high influence, showing that the line computed with this point deleted is quite different than the one with it included. This is also demonstrated by the prediction residual $e_{(i)}$, whose absolute value is also large. The leverage statistic is unchanged from 9.19a, as the x position has not changed.

It is also quite important to note the values for the 10th observation which are not large -- the residual itself (e_i) and the standardized residual (e std). These statistics do not indicate the magnitude of the problem. Therefore residuals plots which use e_i or

e std may not display influential observations as such, because the line has been so drawn near to the outlier that its residual does not appear unusual.

9.5.4 Measures of Serial Correlation

One of the assumptions of regression is that the residuals e_i are independent. Many hydrologic data sets on which regression is performed are actually pairs of time series -- precipitation and flow, flow and concentration, concentration of one constituent versus concentration of another. These series often exhibit serial correlation, the dependence or correlation in time sequence between residuals, violating the assumption of independence (figure 9.10). If the sampling frequency is high enough, serial correlation of the residuals is virtually certain to exist. If serial correlation occurs, the following two problems ensue:

1) The estimates of the regression coefficients are no longer the most efficient estimates possible, though they remain unbiased, and
2) The value of s^2 may seriously underestimate the true σ^2.
   This means that all of the hypothesis tests are wrong (H_0 is rejected too easily) and that confidence and prediction intervals are too narrow.

$y^* = 3.65 + 0.11\ x^*$

n = 10 s = 0.60 $R^2 = 0.21$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	3.648	0.270	13.53	0.000
Slope β_1	0.111	0.075	1.48	0.000

OBS#	e_i	h_i	$e_{(i)}$	e std	e stud	DFFITS	D_i
1	−1.096	0.188	−1.350	−2.042	−2.761	−1.330	0.483
2	−0.166	0.131	−0.192	−0.300	−0.282	−0.109	0.006
3	0.599	0.126	0.687	1.077	1.090	0.415	0.084
4	−0.370	0.122	−0.421	−0.663	−0.638	−0.238	0.030
5	−0.325	0.104	−0.363	−0.576	−0.551	−0.188	0.019
6	0.373	0.104	0.417	0.662	0.637	0.217	0.025
7	0.680	0.101	0.757	1.204	1.245	0.417	0.081
8	0.534	0.100	0.594	0.945	0.938	0.313	0.049
9	0.099	0.101	0.110	0.176	0.165	0.055	0.001
10	−0.329	**0.919**	**−4.117**	−1.955	−2.531	**−8.579**	**21.961**

Table 9.4 Regression statistics for the data of Figure 9.19b

One can search for the presence of serial correlation in two ways. The first is graphical: plotting e_i versus i or a measure of time (figure 9.10b). If there is a tendency for the data to "clump,"

positives follow positives, negatives follow negatives, this may mean there is dependence. The clumping could arise for four different reasons: long-term trend, seasonality, dependence on some other serially correlated variable which was not used in the model, serial dependence of residuals, or some combination of these. Examination of a graph of e_i versus time should help to reveal trend or seasonality if they exist. If there is reason to believe it is trend or seasonality (or both), then steps should be taken to remove these features from the residuals by adding additional explanatory variables. Similarly, if there is an important variable missing from the model, plots of e_i versus this variable should show it, and incorporating this new variable may remove the clumpiness of the residuals. This is particularly likely if this new explanatory variable exhibits serial dependence, seasonality, or trend. The residuals from these new regressions can be plotted again to see what effect this had.

9.5.4.1 Durbin-Watson statistic

There are also statistics for evaluating the dependence of residuals. The standard one is the Durbin Watson statistic (Durbin and Watson, 1951). It is very closely related to a serial correlation coefficient. The statistic is

$$d = \frac{\sum_{i=2}^{n} [e_i - e_{(i-1)}]^2}{\sum_{i=1}^{n} e_i^2}$$

A small value of d is an indication of serial dependence. The H_0 that the e_i are independent is rejected in favor of serial correlation when $d<d_L$ which is tabled in time-series texts. The value of d_L depends on the size of the data set, the number of explanatory variables, and α. However, a low value of d will not give any clue as to its cause. Thus, the graphical approach is vital, and the test is only a check. The Durbin Watson statistic requires data to be evenly spaced in time and with few missing values.

9.5.4.2 Serial correlation coefficient

Serial correlation can also be measured by the correlation coefficient between a data point and its adjacent point. As a linear relationship between pairs of points cannot be assumed, the Kendall's or Spearman's coefficients will provide robust measures of serial dependence. To compute whether this serial dependence is in fact significant,

1) Compute the regression between y and x.
2) Order the resulting residuals by the relevant time or space variable t_1 to t_n.

3) Offset or "lag" the vector of residuals to form a second vector, the lagged residuals. The residuals pairs then consist of (e_i, e_{i-1}) for all i from t_2 to t_n. Figure 9.10c plots one such set of data pairs, illustrating their correlation.
4) Compute Kendall's tau (or Spearman's rho) between the pairs (e_i, e_{i-1}). If the correlation is significant, the residuals are serially correlated.

9.5.4.3 What to do if serial correlation is present

If serial dependence cannot be removed by adding new variables, and one wants to make inferences about parameters, then these three options are available.

1) Sample from the data set. For example, if the data set is quite large and the data are closely spaced in time (say less than a few days apart), then simply discard some of the data in a regular pattern. The dependence that exists is an indication of considerable redundancy in the information, so not a great deal is lost in doing this.
2) Group the data into time periods (e.g., weeks, months) and compute a summary statistic for the period such as a time-weighted mean or median, a volume-weighted mean or median, and then use these summary statistics in the regression. This should only be done when the sampling frequency has remained unchanged over the entire period of analysis.
3) Use much more sophisticated estimation methods, specifically Box and Jenkins (1976) transfer function models, or regression with autoregressive errors Johnston (1984).

9.6 Transformations of the Response (y) Variable

The primary reason to transform the response variable is because the data are heteroscedastic -- the variance of the residuals is a function of x. This situation is very common in hydrology. For example, suppose a rating curve between stage (x) and discharge (y) at a stream gage has a standard error of 10 percent. This means that whatever the estimated discharge, the standard error is 10 percent of that value. The absolute magnitude of the variance around the regression line between discharge and stage therefore increases as estimated discharge increases. The ideal variance stabilizing transformation in these cases is the logarithm because a multiplicative relationship, such as standard error = 0.10•estimate, becomes a constant additive relationship after log transformation. This satisfies the regression assumptions. The two topics that require careful attention when transforming y are:

1) deciding if the transformation is appropriate, and
2) interpreting resulting estimates.

9.6.1 To Transform or Not to Transform?

The decision to transform y should generally be based on graphs. First develop the best possible non-transformed model. This should entail considering all sorts of transformations of x (or

multiple x variables) to get a good and reasonable fit. Then plot e_i vs. $\hat{y}_i$ to check for heteroscedasticity, do a probability plot for e_i to check for normality, and examine the function for unreasonable results (i.e., predictions of negative values for variables that can't go negative). If serious problems arise for any of these reasons, transform y and repeat the process. If both the transformed and untransformed scales have problems, then either look for a different transformation or accept the lesser of two evils.

Two methods are available to numerically judge whether or not to transform y. The first is to perform a series of transformations, perform regressions, and choose the transformation which maximizes the probability plot correlation coefficient (PPCC) for the regression residuals. This optimizes the normality of residuals. The second method is similar, optimizing for linearity. It searches for the minimum sum of squared errors SSE from a series of regressions using transformed and scaled y variables (Montgomery and Peck, 1982, p.94). The transformations used are scaled versions of the ladder of powers called "Box-Cox transformations". Scaling is required in order to compare the errors among models with differing units of y. Either numerical method can be a useful guide to selecting several candidate transformations from which to choose. However, the final choice should be made only after looking at residuals plots.

The key thing to note here is that **comparisons of R^2, s, or F statistics between transformed and untransformed models cannot easily be used to choose among them**. Each model is attempting to predict a different variable (y, log(y), 1/y, etc.). The above statistics therefore measure how well different variables are predicted, and so cannot be directly compared. Instead, the appropriate response variable is one which fits the assumptions of regression well -- linear and homoscedastic, having a good residuals plot. Once a hydrologist has developed some experience with certain kinds of data sets, it is quite reasonable to go directly to the appropriate transformation without a lot of investigation. One helpful generalization is that any y variable that covers more than an order of magnitude of values in the data set, as sediment discharge or bacterial densities typically do, probably needs to be transformed.

9.6.2 Consequences of Transformation of y

Let's take a particular, but rather common, case of a transformed regression problem. The model is

$$\ln(L) = \beta_0 + \beta_1 \ln Q + \varepsilon$$

where ln is the natural log, L is constituent load (tons/day), and Q is discharge (cubic feet per second). Let us further assume that the ε values are normal with mean zero and variance σ^2.

Figure 9.20 illustrates a data set typical of such L vs. Q data, shown here as a log-log plot. The lines results from a SLR done in log units. The middle line is the regression line and the 50% and 95% prediction intervals are shown. Note that, because of the normality assumption, the

prediction intervals are symmetric about the regression line. For any given Q value the five lines on the graph represent five different percentage points on the conditional distribution of ln(L). They are the 2.5, 25, 50 (median), 75, and 97.5 percentage points. The median also happens to be the conditional mean for ln(L) because when normality is assumed the median = mean. So the regression line falls on both the conditional median and mean value for ln(L).

Figure 9.21 takes each of these data points and lines and replots them in the original units (L versus Q). The five curves remain the 2.5, 25, 50, 75, and 97.5 percentage points on the conditional distribution. Now however this distribution of L conditional on Q is lognormal, not a normal distribution. Note the asymmetry of the curves around the regression line. For a lognormal distribution the mean is not equal to the median. While the central line remains the conditional median following transformation, the conditional mean of L will always lie somewhere above the regression line.

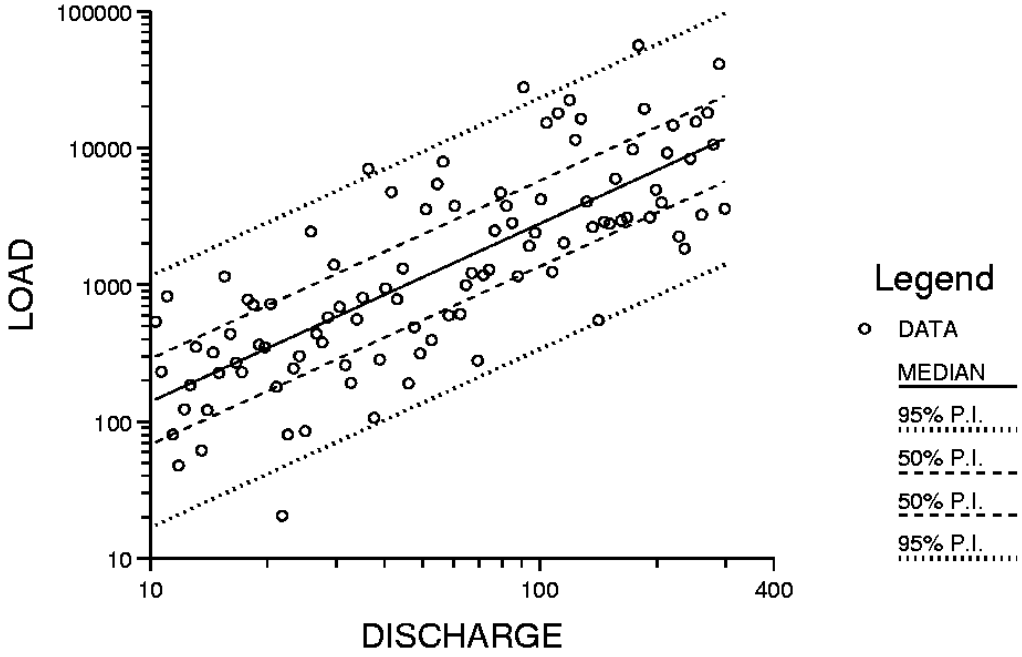


Figure 9.20 Prediction intervals and log-log regression in log units.

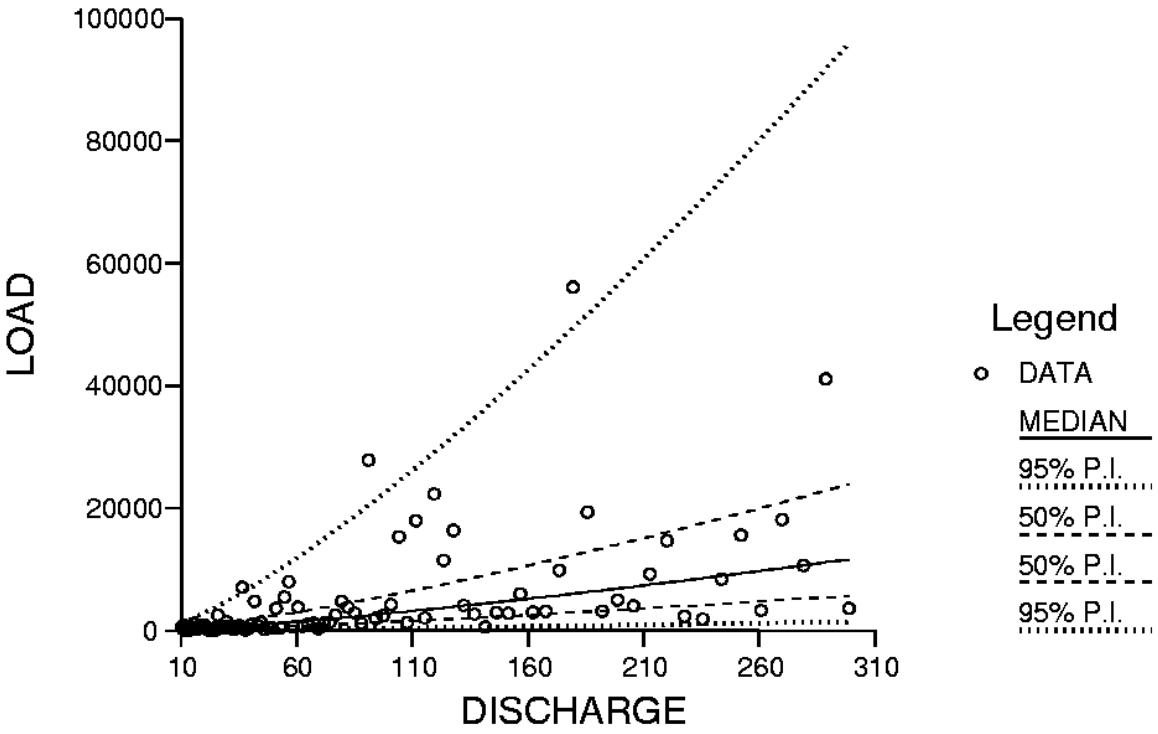


Figure 9.21 Prediction intervals and log-log regression re-expressed in original units.

9.6.3 Computing Predictions of Mass (Load)

9.6.3.1 Median or "rating curve" estimate of mass

When the objective is estimating the mass of sediment (or nutrient or contaminant) entering a lake, reservoir, or estuary, the mean for each of many short time periods can be estimated by regression and summed to estimate the total (or mean) mass over a longer time period. This is appropriate because the sum of the means equals the mean of the sum. However, simply transforming estimates from a log-regression equation back into the original units for y provides a median estimate of L, not a mean. Unfortuantely, this has been the traditionally-used method since Miller (1951). The sum of these medians provides an estimate of the mean of L which is biased low. As the sum of the medians is not the median of the sum, it is difficult to state what the sum of these median values represents, except that it underestimates the long-term mean load.

Ferguson (1986) points out for some very realistic cases that using the median or rating curve estimate for loads:

$$\hat{L}_m = \exp [b_0 + b_1 \ln(Q_0)]$$

will result in underestimates of the mean by as much as 50%. The question then is how to compensate for this bias. The following two methods, one assuming a normal distribution of the logs and the other a nonparametric method, attempt to correct for this bias of the median estimate.

9.6.3.2 Parametric or "MLE" estimate of mass

If the residuals in natural log units were known to be normal and the parameters of the model (β_0, β_1, σ^2) were known without error, the theory of the lognormal distribution (Aitchison and Brown, 1981) provides the following results:

$$\text{Median of L given } Q_0 = \exp[\beta_0 + \beta_1 \ln(Q_0)] = L_m = \exp[\beta_0] \cdot Q_0^{\beta_1}$$

$$\text{Mean of L given } Q_0 = E[L|Q_0] = \exp[\beta_0 + \beta_1 \ln(Q_0) + 0.5\,\sigma^2] = L_m \cdot \exp[0.5\,\sigma^2]$$

$$\text{Variance of L given } Q_0 = V[L|Q_0] = [L_m \cdot \exp(0.5\,\sigma^2)]^2 \cdot [\exp(\sigma^2) - 1]$$

These equations would differ if base 10 logarithms were used (Ferguson, 1986).

Unfortunately the true population values β_0, β_1, and σ^2 are never known in practice. All that is available are the estimates b_0, b_1, and s^2. Ferguson (1986) assumed these estimates were the true values for the parameters. His estimate of the mean is then

$$\hat{L}_{MLE} = \exp[b_0 + b_1 \ln(Q_0) + 0.5\,s^2]$$

When n is large (>30) and σ is small (<0.5), $\hat{L}_{MLE}$ is a very good approximation. However, when n is small or σ is large, it can <u>overestimate</u> the true mean -- it overcompensates for the bias. There is an exact unbiased solution to this problem which was developed by Bradu and Mundlak (1970). It is not given here due to the complexity of the formula. Its properties are discussed in Cohn (1988). Even so, the validity of Bradu and Mundlak's solution depends on the normality of the residuals which can never be assured in practice.

9.6.3.3 Nonparametric or "smearing" estimate of mass

There is an alternative approach which only requires the assumption that the residuals are independent and homoscedastic. They may follow any distribution. This is the "smearing" estimate of Duan (1983). In the case of the log transform it is

$$\hat{L}_D = \exp[b_0 + b_1 \ln(Q_0)] \cdot \frac{\sum_{i=1}^{n} \exp[e_i]}{n}$$

The smearing estimator is based on each of the residuals being equally likely, and "smears" their magnitudes in the original units across the range of x. This is done by re-expressing the residuals from the log-log equation into the original units, and computing their mean. This mean is the "bias-correction factor" to be multiplied by the median estimate for all x_0. Even when the residuals in log units are normal, the smearing estimate performs very nearly as well as Bradu and Mundlak's unbiased estimator. It avoids the overcompensation of Ferguson's approach. As it is robust to the distribution of residuals, it is the most generally-applicable approach.
The smearing estimator can also be generalized to any transformation. If Y = f(y) where y is the response variable in its original units and f is the transformation function (e.g., square root, inverse, or log), then

$$\hat{y}_D = \frac{\sum_{i=1}^{n} f^{-1}(b_0 + b_1 X_0 + e_i)}{n}$$

where b_0 and b_1 are the coefficients of the fitted regression and e_i are the residuals ($Y_i = b_0 + b_1 X_0 + e_i$) , f^{-1} is the inverse of the selected transformation (e.g., square, inverse, or exponential, respectively) and X_0 is the specific value of X for which we want to estimate y.

9.6.4 An Example

Total phosphorus loads are to be estimated for the Illinois River at Marseilles, Illinois, drainage area 8259 square miles, for the period 1972-1985. The data are contained in Appendix C10. The 96 measurements of load are plotted in figure 9.22 as a function of discharge. As loads were not sampled for each day during this time period, estimates of load for unsampled days are to be obtained from a regression equation as a function of discharge.

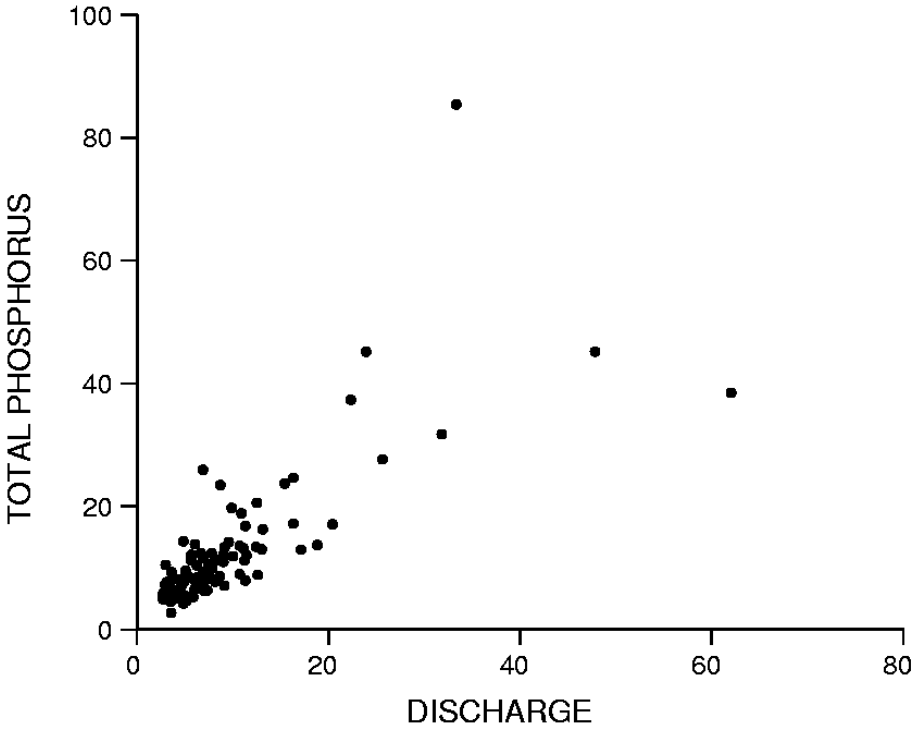


Figure 9.22 Total phosphorous load and stream discharge for the Illinois River

The first question is whether a log transform of load is necessary to develop a good prediction equation. From figure 9.22, the variance appears to greatly increase as discharge increases. Therefore a log transformation of phosphorus is attempted. This results in a curvilinear pattern, so the log of discharge is computed and used as the explanatory variable. As seen in figure 9.23, the transformation of both load and discharge results in a linear, homoscedastic relationship. A residuals plot in figure 9.24 shows little evidence of structure, indicating that the units are appropriate. Therefore these units are used for the regression. Table 9.5 gives the relevant regression statistics.

$\ln(L) = 0.80 + 0.76 \ln(Q)$

n = 96 s = 0.339 $R^2 = 0.68$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	0.799	0.114	7.03	0.000
Slope β_1	0.761	0.054	14.10	0.000

Table 9.5 Regression statistics for the Illinois River phosphorus data

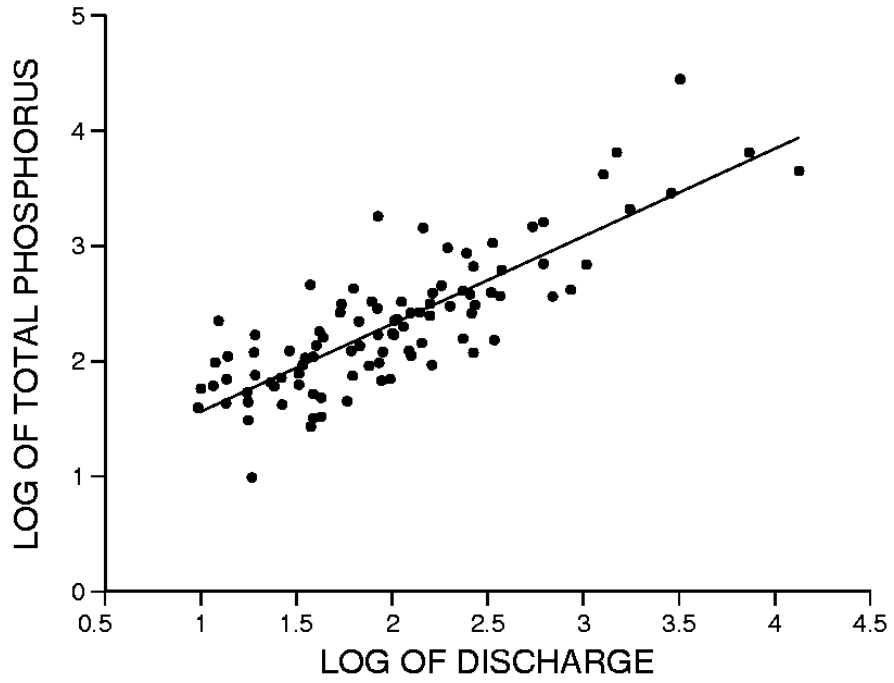


Figure 9.23 Log-log relation between phosphorous and discharge for the Illinois River

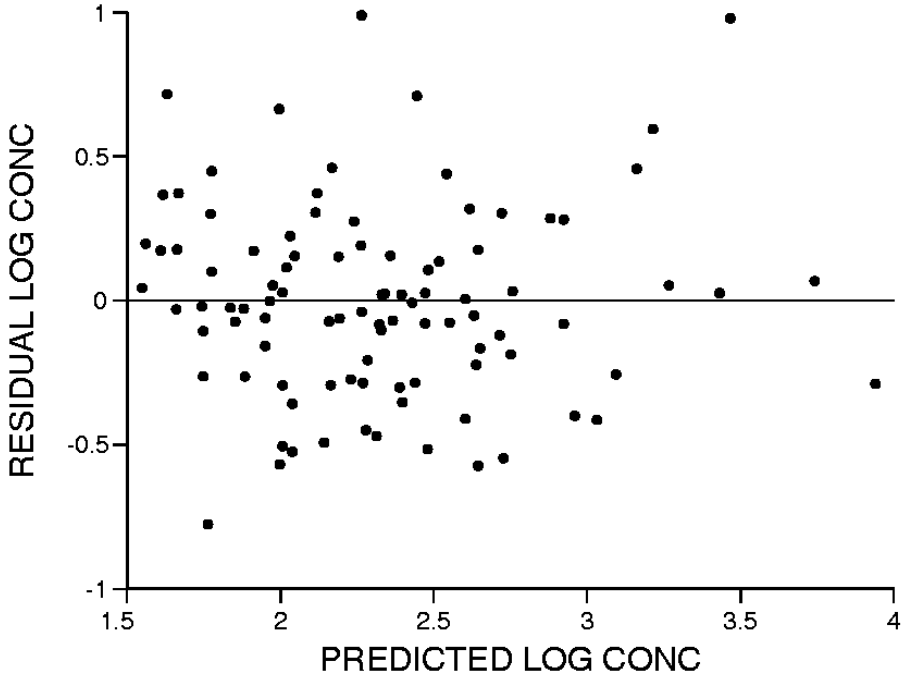


Figure 9.24 Residuals plot for ln(phosphorous) versus ln(discharge)

To illustrate the bias in phosphorus loads for the rating curve method, and the bias correction capabilities of the other two methods, estimates of all three will be computed here for the 96

days for which data exist. These values can then be compared to the "true" loads computed from the observed data.

The results from this regression are these	Mean Load	Error
true	= 12.64	--
median estimate	= 11.72	-7.3%
MLE estimate	= 12.41	-1.8%
smearing estimate	= 12.44	-1.6%

The median estimate is biased low, while the MLE and smearing estimates are close to each other and to the true value (figure 9.25). The MLE and smearing estimates should be expected to be similar here, as the residuals are fairly symmetric, n is large and s is small. These are the conditions under which the MLE works well. Had s been large (>1) or n small (<30) the MLE would probably have had a positive bias, and only the smearing estimate would have come close to the true value.

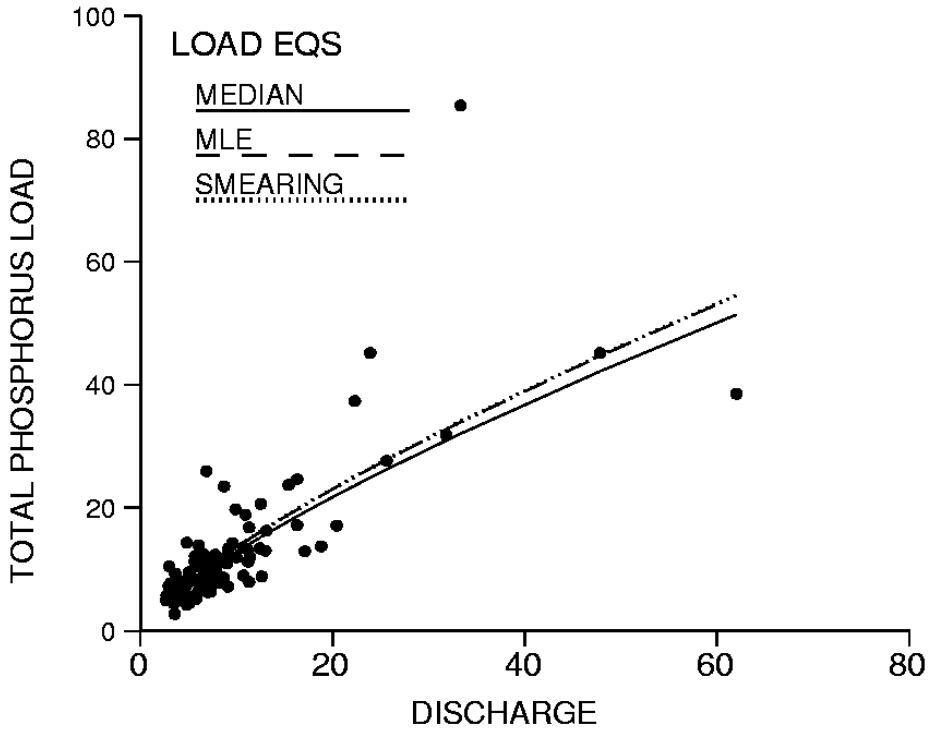


Figure 9.25 Load estimate curves with and without bias correction for Illinois R. data

9.7 Summary Guide to a Good SLR Model

1) "Should x be transformed, and if so, how?" Considerable help can come from a statistic such as R^2 (maximize it), or s (minimize it), but these numbers alone do not insure a good model. Many transformations can be rapidly checked with such statistics, but always look at a residual versus predicted plot before making a final decision. Transform x if the residuals plot appears non-linear but constant in variance, always striving for a linear relation between y and x.

2) "Should y be transformed, and if so, how?" Visually compare the transformed-y model to the untransformed-y model using their residuals plots (residual versus predicted). The better model will be more:
   1) linear,
   2) homoscedastic, and
   3) normal in its residuals.

   The statistics R^2, s, t-statistics on β_1, etc. will not provide correct information for deciding if a transformation of y is required.

   Should estimates of mass (loads) be developed using an equation having transformed-y units, the transformation bias inherent in the process must be compensated for by use of the smearing estimate, or MLE estimate when appropriate.

When there are multiple explanatory variables, more guidelines are required to choose between the many possible combinations of adding, deleting and transforming the various x variables. These are discussed in Chapter 11.

Exercises

9.1 Bedinger (1961) graphically related median grain size of alluvial aquifer materials in the Arkansas River Valley to their yield, in gallons per day per square foot. This enabled estimates of yield to be made at other locations based on measured grain-size analyses. Compute a regression equation to predict yield, based on the data in Appendix C11.

9.2 Estimate the mean yield in gallons per day per square foot available from four wells which together compose the public supply of a small town in the Arkansas River Valley. The wells have screens with identical cross-sectional areas. Median grain sizes for the units they draw from are: 0.1, 0.2, 0.4 and 0.6 millimeters.

9.3 Find a transformation of discharge for the Cuyahoga River TDS example which might improve on the $\log_{10}$ transformation used throughout the chapter. The data are found in Appendix C9. Obvious candidates include the ladder of power transformations. Another class of transformations that has been shown to work well for surface-water chemistry is the hyperbolic transformations (see Johnson, et al., 1969). The form of this transformation is $x=1/(1+kQ)$ where k is some constant supplied by the hydrologist. Some general advice about selecting k is that it's not worth the effort to try and get it "right" to a precision better than about half an order of magnitude. A good range to work in is

$$1/(100(\overline{Q}) < k < 100/(\overline{Q})$$

where $(\overline{Q})$ is the mean discharge.

The questions you should answer are:

a) What is a good transformation of Q to use in estimating TDS? (There is no "best" transformation, but there are several good ones.)
b) Describe your preferred model and indicate some reasons you might be concerned about it and might want to take steps to "fix" it in some fashion. (You will get a chance to later.)
c) What does it tell you about TDS behavior in the Cuyahoga River?
d) A question for the mathematically inclined. If k is set to some very large value (say around $100/\overline{Q}$), what other model does the hyperbolic approximate? If b is set to some very small value (say around $1/100\overline{Q}$), what other model does it approximate?

9.4 Objections have been raised to regressions such as load (L) versus stream discharge (Q) because Q is used to calculate L. This "spurious correlation" between Q and L can be avoided by using concentration (C) instead of load as the dependent variable. Loads would then be predicted from the estimated C. What do you think? How will the results using C compare to those using L as the regression's response variable? To answer this, perform the regression for the Illinois phosphorus data of section 9.6.4 and produce the 96 load estimates using ln(C in mg/L) instead of ln(L in tons per day). The data are found in Appendix C10. Note that the units of Q (thousands of cfs) mean that L = 2.7 Q•C. What happens to the regression coefficients and the associated statistics such as R^2, s, t-ratios, etc., when ln(C) rather than ln(L) is used? What is the appropriate conclusion to this controversy?